PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/29, 15/31, C07K 14/195, 14/415, C12Q 1/68, G01N 33/50

A2

(11) International Publication Number: **WO 99/54470**

(43) International Publication Date: 28 October 1999 (28.10.99)

(21) International Application Number: PCT/EP99/02635

(22) International Filing Date: 20 April 1999 (20.04.99)

(30) Priority Data:
9808423.9 22 April 1998 (22.04.98) GB

(71) Applicant *(for all designated States except US)*: GLAXO GROUP LIMITED [GB/GB]; Glaxo Wellcome House, Berkeley Avenue, Greenford, Middlesex, UB6 0NN (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ARIGONI, Fabrizio [CH/CH]; 2 rue Maurice, CH–1204 Geneva (CH). EDGERTON, Michael, David [US/US]; Dekalb Genetics, 62 Maritime Drive, Mystic, CT 06355 (US). LOFERER, Hannes [AT/DE]; Alpenstrasse 74, D–82538 Geretsried (DE). PEITSCH, Manuel, C. [CH/CH]; En Jaquerod, CH–1855 La Forclaz (CH).

(74) Agent: LEAROYD, Stephanie, Anne; Glaxo Wellcome plc, Glaxo Wellcome House, Berkeley Avenue, Greenford, Middlesex UB6 0NN (GB).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: BACTERIAL YGJD POLYPEPTIDE FAMILY

(57) Abstract

This invention relates to a family of bacterial polypeptides which are considered essential for growth of both gram negative and gram positive bacteria. The family has been identified by a number of methods including computer based algorithms. The use of such polypeptides and the genes which encode them as tools for identifying novel broad spectrum antibiotics is described.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

BACTERIAL YGJD POLYPEPTIDE FAMILY

This invention relates to a family of bacterial polypeptides which are required for growth of both gram negative and gram positive bacteria, the genes which encode them and the use of such polypeptides and genes as tools for identifying novel broad spectrum antibiotics.

New antibiotics are urgently needed in current medical practice as both serious bacterial infections and multiply antibiotic resistant strains are becoming increasingly prevalent (Proc. Natl. Acad. Sci USA (1994) 91:2420-2427; New England J. Med. (1994) 330:1247-1251). The increase in number of serious infections has been ascribed to a variety of causes, including: 1) Increasing age of the general population, 2) increasingly long and complex surgeries and 3) a growing immuno-suppressed population associated with cancer therapies, organ transplants and HIV infection. Overuse of antibiotics in both medical and agricultural settings, improper sanitation and a general lack of concern about antibiotic resistant organisms have all contributed to the increasing frequency of multiply antibiotic resistant bacteria. Taken together, these two trends suggest that we will soon be faced with bacterial infections which are resistant to all therapies. Indeed, the first report of vancomycin-resistant *S. aureus* has just been published (Lancet (1997) 350:1670-1673).

Identification of conserved essential proteins is a key step in the development of broad-spectrum antibiotics. If a target protein is conserved across taxonomic lines, the possibility that antibiotics acting on that protein will be effective on a wide range of bacteria is maximized. As examples, DNA gyrase and RNA polymerase are found in all bacteria, which helps to explain why quinolones and rifampicin are good broad-spectrum antibiotics. However, not all bacteria synthesize peptidoglycan, which explains why b-lactam antibiotics are ineffective against *Chlamydia*, *Rickettesia* and *Legionella* species. The recent publication of several complete eubacterial genomic sequences (Science (1995) 270:397-403; Science (1997) 277: 1453-1474; Nature (1997) 390:249-256) allows the identification of bacterial proteins which have orthologues in all of the sequenced genomes. This approach has lead to the identification of many conserved protein families (Science (1997) 278:631-637). In some cases a biochemical function for the conserved family may

be deduced from their predicted amino acid sequence. In other cases no function can be predicted for the protein family. However, it is impossible to predict the physiological role of a protein or protein family without detailed characterisation of at least one family member.

Following identification of a conserved bacterial protein family, the protein must be shown to be essential for bacterial viability if it is to serve as an antibiotic target. Genetic systems have been developed to demonstrate a genes essentiality in both *E. coli* (J. Bacteriol. (1997) 179:6228-6237) and *B. subtilis* (Genes Dev. (1991) 177:4194-4197). In some instances these systems suffer either from a reliance on negative data, failure to disrupt a given gene, or insufficient repression of the candidate gene, which can lead to misidentification of genes essentiality. Clean data from taxonomically diverse bacteria, such as gram negative and gram positive strains offers the best evidence that a conserved bacterial protein family is essential for viability and will make a good broad-spectrum antibiotic target.

We have identified a family of conserved bacterial genes which we have designated the ygjD gene family, after the name given to the *E. coli* gene family member. These genes have not been previously isolated nor the polypeptides expressed as no function has been ascribed to these genes. It has now been discovered that this family of genes encodes a family of polypeptides which are essential for the survival their host bacteria.

The invention therefore provides an isolated polypeptide of the ygjD family as defined below particularly for use in the identification of novel antibiotic agents. The polypeptides of the present invention are believed to be essential to the viability of a wide range of bacteria including both gram positive and gram negative bacteria.

Any one of the following three methods may be used to identify members of the ygjD family as claimed herein;

BLAST searches (J. Mol. Biol. (1990) 215:403 -10 and Meth. Enzymol. (1996) 266: 131-141, 227-258 both incorporated herein by reference) may be carried out using the ygjD family member sequences as described in Figure 1. Such searches involve using in succession as query sequences, each of the existing ygjD protein

family member sequences to identify other full length members of the ygjD family of proteins. Such family members yield high-scoring segment pairs (HSP) scores of greater than 100 in comparison to at least one member of the ygjD family when the BLAST algorithm described in the reference above is used with a particular scoring matrix (a BLOSUM62 matrix - Proteins (1993) 17:49-61 incorporated herein by reference).

Profile based searches (Proceedings of the second International Conference on Intelligent Systems for Molecular Biology, pp28-36, AAAI Press, Menlo Park California, 1994 incorporated herein by reference) may be carried out using position-dependent scoring matrices defined for the ygjD family members. These searches use a table compiled from a multiple sequence alignment which describes distinctive sequences of amino acids as probability values for each residue at each position in the gene family to identify other proteins which contain similar sequences of amino acids.

Motif based searches (Nucleic Acids Res. (1995) 24:189-196 incorporated herein by reference) may be carried out using PROSITE patterns defined for the ygjD family members. These searches involve the representation as patterns, of the conserved sequence elements identified in the profile searches.

The isolated polypeptides of the invention may therefore be characterised by:

i) an HSP score of greater than or equal to 100 when compared with one of the sequences of Figure 1 when the BLAST algorithm is used with a LOSUM62 scoring matrix; or

ii) containing a set of amino acid sequences which are positively identified when position dependent scoring matrices according to Tables 1-4 are used with MAST to yield a p-value of less than $1x10^{-50}$; or

iii) comprising at least one of the following amino acid sequences:

[LIV](2)-[SCT]-G-G-H-X(17,21)-D-D-[AST]-X-G-E-X(2)-D-K;
A-X(3)-P-G-L-X(3)-l-X(2)-G-X(13)-P-X(5)-H-X(3)-H

[VIL]-I-[GSAT]-[VILFM]-E-[TS]-[TS]-C-D-[DE]; and
G-[LIV]-V-P-E-[LIV]-A-[AST]-R-X-H;

wherein
the letters denote an amino acid in one letter code,
the square brackets denote a single amino acid,
the amino acids within the square brackets are alternatives,
X is any one amino acid residue, and
the numbers in the curved brackets refer to the number of residues at that position;

or

iv) [KR]-[GSAT]-X(4)-[FYWLH]-[DQNGK]-X-P-X-[LIVMFY]-X(3)-H-X(2)-[AG]-H-[LIVM]

wherein
the letters denote an amino acid in one letter code,
the square brackets denote a single amino acid,
the amino acids within the square brackets are alternatives,
X is any one amino acid residue, and
the numbers in the curved brackets refer to the number of residues at that position.

In a preferred aspect of the invention all three of the amino acid sequences listed under iii) are present.

The invention also provides an isolated polypeptide sequence as set out in any of Figures 2a-d.

The polypeptides are preferably recombinant and ideally purified to homogeneity.

Also included as polypeptides according to the invention are variants, analogues and derivatives. Particularly those in which a number of amino acids have been substituted, deleted or added. Polypeptides which have at least 70% identity to any

of the polypeptide sequences according to the invention, in particular the sequences of Figures 2a-d are encompassed within the invention. Preferably the identity is at least 80%, more preferably at least 90% and still more preferably at least or greater than 95% identity for example 97%, 98% or even 99% identity to any of the sequences according to the invention, in particular the sequences of Figures 2a-d.

Such polypeptides may also be fragments. In this regard a fragment is a part of a polypeptide according to the invention which retains sufficient identity of the original polypeptide to be effective for example in a screen. Such fragments may be fused to other amino acids or polypeptides or may be comprised within a larger polypeptide. Such a fragment may be comprised within a precursor polypeptide designed for expression in a host. Therefore in one aspect the term fragment means a portion or portions of a fusion polypeptide or polypeptide derived from a polypeptide according to the invention.

Fragments also include portions of a polypeptide according to the invention characterised by structural or functional attributes of a polypeptide according to the invention. These may have similar or improved chemical or biological activity or reduced side-effect activity. For example fragments may comprise an alpha helix or alpha-helix forming region, beta sheet and beta-sheet forming region, turn and turn forming regions, coil and coil-forming regions, hydrophilic regions, hydrophobic regions, amphipathic regions (alpha or beta), flexible regions, surface-forming regions, substrate binding regions and regions of high antigenic index.

Fragments or portions may be used for producing the corresponding full length polypeptide by peptide synthesis.

Specific polypeptides according to the invention include the polypeptides of *Borrella burgdorferi, Treponema pallidium, Synechocystis* sp. Strain PCC6803, *Helicobacter pylori, Arabidopsis thaliana, Haemophilus influenza, Mycobacterium tuberculosis, Mycobacterium leprae, Pasturella haemolytica, Mycoplasma genitalium, Mycoplasma pneumoniae, Streptococcus pneumoniae, Streptococcus pyogenes, Bacillus subtilis* and *Escherichia coli*.

The present invention further provides isolated polynucleotides which encode the polypeptides as defined herein, polynucleotides complementary thereto, or polynucleotides hybridising to any of the aforesaid polynucleotides. Isolated polynucleotides have been removed by separation from their natural environment and those materials with which they are naturally associated. Preferably these polynucleotide molecules are provided in recombinant form (i.e. combined with one or more heterologous sequences).

Polynucleotide molecules which hybridise to polynucleotides encoding substances of the present invention, or to complementary polynucleotides thereto, preferably do so under stringent hybridisation conditions. One example of stringent hybridisation conditions which is sometimes used is where attempted hybridisation is carried out at a temperature of from about 35°C to about 65°C using a salt solution which is about 0.9 molar. However, the skilled person will be able to vary such conditions as appropriate in order to take into account variables such as probe length, base composition, type of ions present, etc.

The invention also provides polynucleotide variants, analogues, derivatives and fragments which encode polypeptides according to the invention. Polynucleotides are included which preferably have at least 70% identity over their entire length to a polynucleotide encoding a polypeptide according to the invention, most preferably those set out in Figures 2a-d. More preferred are those sequences which have at least 80% identity over their entire length to a polynucleotide encoding a polypeptide according to the invention. Even more preferred are polynucleotides which demonstrate at least 90% for example 95%, 97%, 98% or 99% identity over their entire length to a polynucleotide encoding a polypeptide according to the invention.

Polynucleotide molecules of the present invention may be used as probes for other members of the gene family or in anti-sense therapy to block or to reduce the expression of one or more of the polypeptides of the invention. Since these substances are believed to be essential to the bacteria expressing them, blocking or reducing their expression can provide an effective way of treating bacterial mediated diseases or disorders. Polynucleotides may also be used directly in screening and in generating whole cell screens by expression of a polypeptide of the inventions.

As part of the isolation process or thereafter the polynucleotides may be joined to other polynucleotides such as to form fusions or to regulatory elements for expression. Isolated polynucleotides alone or joined to other polynucleotides can be in introduced into a vector which itself will contain other elements of DNA or RNA for expression in a host cells. The invention therefore comprises a vector containing a polynucleotide generally operatively linked to appropriate expression control sequences.

Vectors for use in the invention include plasmid vectors, phage vectors and DNA or RNA viral vectors. These vectors may include gene sequences which render them inducible under certain conditions such as manipulation of the environmental conditions under which the host cells are maintained for example by temperature alteration or nutrient additives. Regulatory sequences include for example a promoter to direct mRNA transcription. Such promoters include for example *E. coli.* lac, trp, tac and araBAD as well as the SV40 early and late promoters Such systems and sequences would be well known to those skilled in the art.

Host cells expressing a polynucleotide of the present invention can be generated by any of the traditional routes such as transfection or electroporation see for example Davis et al, Basic Methods in Molecular Biology, (1986) and Sambrook et al Molecular Cloning: A Laboratory Manual, 2nd Edition., Cold Spring Harbor Lab. Press, Cold Spring Harbor, N.Y. (1989).

This invention also provides a method for identification of molecules such as antagonists, that bind to the polypeptide or a polynucleotide encoding a polypeptide of the present invention.

Selective whole-cell screens combine the sensitivity and specificity of *in vitro* biochemical assays with the direct demonstration of *in vivo* activity seen in whole cell screens. Biochemical assays for inhibition of polypeptide activity with purified polypeptides or bacterial extracts can be more sensitive than whole cell killing assays and provide direct evidence for a compound's mode of action. However, this approach requires that the target polypeptide is known and the activity of the polypeptide be amenable to *in vitro* assays. Nor does it address other factors, such

as membrane permeability or compound stability, which can limit a compounds effectiveness as an antibiotic.

Whole cell screening of compounds for killing activity will identify molecules which kill cells at the concentrations tested, but provide no information on the mode of action of the compound and may not have the sensitivity needed to detect less potent compounds. Bacterial strains which contain surrogate markers whose activity is linked to that of the target gene or which have been engineered to over-express or under-express the target polypeptide can be used for selective whole-cell screens.

Surrogate markers, easily assayed reporter molecules whose activity is tightly coupled to the activity of the polypeptide being studied, may be used as a means of assaying antibiotics. The invention further provides a host cell comprising a vector as defined herein and a reporter gene encoding a reporter molecule whose activity is linked to that of the polypeptide encoded by the vector. Examples of such systems include a transcriptional fusion of the *E. coli* lacZ gene to vanH promoter in a *B. subtilis* strain expressing VanS and R as a reporter for inhibition of cell wall biosynthesis (J. Bacteriol. (1996) 178:6305-6309), the use of lacZ transcriptional and translational fusions to rpoB and rpoC to monitor RNA polymerase activity (Mol. Microbiol. (1996) 19:483-493) and the use of a secA-lacZ gene fusion as a reporter for inhibition of secA activity (Genetics (1988) 118:571-579).

When the function of a gene is unknown, surrogate markers for the activity of the gene can be identified using at least two approaches. Two dimensional electrophoresis coupled with mass spectrometry analysis of isolated polypeptides, proteome mapping, has been used to identify specific polypeptides which increase in abundance in response to polypeptide or RNA synthesis inhibitors (Microbial & Comparative Genomics (1996) 1:375). Tightly regulated promoters used to demonstrate that the *E. coli* and *B. subtilis* conserved, essential polypeptides are essential can also be used to reduce the concentrations of these polypeptides. In a manner similar to that described above, proteome maps generated from bacteria depleted of the conserved essential genes can be used to detect polypeptides which change in abundance as compared to wild-type bacteria. Transcriptional or translational fusions to these polypeptides can be used as reporter molecules to screen for antagonists of members of the conserved essential gene family. As an

alternative to proteome mapping, transposons or other mobile genetic elements containing reporter genes can be used to search for reporter molecules. Such an approach has been used to identify vancomycin responsive genes in *S. aureus* (Antibiot. (Tokyo) (1991) 44:210-217). As with proteome mapping, bacteria in which conserved essential genes are controlled by tightly regulated promoters can be used to screen for transposon carrying strains in which expression of promoterless reporter genes is induced upon depletion of the polypeptides.

Once a reporter gene has been identified, screening of compounds for induction or inhibition of the marker can be undertaken. Standard broth or plate assays can be used in many different formats. Such assays will detect molecules which antagonise the response which couples the activity of the conserved, target polypeptide to the reporter molecule. Thus, the compounds identified may act directly upon the target polypeptide or on another stage in the pathway which leads to activation of the reporter.

Screens for inhibitors of the target which do not require the use of surrogate markers may be designed by manipulating expression levels of the target polypeptide. For example, quinolone resistant strains of *E. coli* have been made by over-expression of gyrA (FEMS Microbiol. Lett. (1997) 154:271-276), over-expression of alanine racemase has been shown to increase resistance to cycloserine in *M. smegmatis* (J. Bacteriol. (1997) 179:5046-5055), and multicopy plasmids carrying murZ have been shown to increase phosphomycin resistance in both *E. coli* (J. Bacteriol. (1992) 174:5748-5752) and *A. calcoaceticus* (FEMS Microbiol. Lett. (1994) 117:137-142). Similarly, strains more sensitive to antibiotics may be made by reducing expression levels of the polypeptide targeted by the antibiotic. Over or under-expression of members of the conserved, essential gene family may be used to screen for antibiotics which act either directly on gene or gene product or indirectly on the pathway which it is involved.

Another example of an assay for antagonists is a competitive assay that combines the polypeptide of the present invention and a potential antagonist with membrane-bound binding molecules, recombinant binding molecules, natural substrates or ligands, or substrate or ligand mimetics, under appropriate conditions for a competitive inhibition assay. The polypeptide can be labelled, such as by

radioactivity or a colorimetric compound, such that the number of polypeptide molecules bound to a binding molecule or converted to product can be determined accurately to assess the effectiveness of the potential antagonist.

The present invention therefore provides a method of assaying compounds for activity against bacteria comprising:

i) providing a polypeptide according to the invention;
ii) contacting said polypeptide with candidate inhibitory compounds; and
iii) measuring for binding to said polypeptide or fragment.

The present invention also provides a method of assaying compounds for activity against bacteria comprising:

i) expressing a polypeptide according to the invention in a host cell;
ii) contacting said cell with candidate inhibitory compounds; and
iii) measuring cell death.

The present invention further provides a method of screening for an antibiotic which method comprises:

i) transfecting a host cell with a vector comprising a polynucleotide encoding a polypeptide as defined herein;
ii) allowing the host cell to express the polynucleotide;
iii) increasing the level of expression of the polypeptide as defined herein; and
iv) assaying for increased resistance.

Alternatively the method may be carried out as above but the level of expression of the polypeptide is decreased and the cells are assayed for increased sensitivity to an inhibitor.

The present invention also provides a method of assaying compounds for activity against bacteria comprising:

i) generating a bacterial strain containing a reporter gene linked to the gene encoding a polypeptide according to the invention;
ii) contacting said strain with candidate inhibitory compounds; and
iii) measuring for induction or inhibition of said marker.

Potential antagonists include small organic molecules, ions which interact specifically with a polypeptide or polynucleotide for example a substrate, cell membrane component, receptor a fragment thereof or a peptide. Such molecules may include antibodies, antibody-derived reagents or chimaeric molecules.

Potential antagonists also may be small organic molecules, a peptide, a polypeptide such as a closely related protein or antibody that binds to the same sites on a binding molecule without inducing functional activity of the polypeptide of the invention.

The antibodies may be monoclonal or polyclonal. Techniques for producing monoclonal and polyclonal antibodies which bind to a particular polypeptide are now well developed in the art. They are discussed in standard immunology textbooks, for example in Roitt *et al* (*Immunology*, Churchill Livingston, 2nd Edition (1989)).

In addition to whole antibodies, the present invention covers variants thereof which are capable of binding to an epitope present or a substance of the present invention. The variants may be antibody fragments or synthetic constructs. Examples of antibody fragments and synthetic constructs are given by Dougall *et al* in *Tibtech* 12 372-379 (September 1994). Antibody fragments include Fab and Fv fragments.

Other synthetic constructs include CDR peptides. These are synthetic peptides comprising antigen binding determinants. Peptide mimetics may also be used. These molecules are usually conformationally restricted organic rings which mimic the structure of a CDR loop and which include antigen-interactive side chains. Synthetic constructs include chimaeric molecules. Thus, for example, humanised antibodies or derivatives thereof are within the scope of the present invention. An example of a humanised antibody is an antibody having human framework regions, but a rodent or other non-human hypervariable regions. Synthetic constructs also include molecules comprising a covalently linked moiety which provides the molecule with some

desirable property in addition to antigen binding. For example the moiety may be a label (e.g. a fluorescent or radioactive label) or a pharmaceutically active agent.

Other potential antagonists include antisense molecules (see Okano, J. Neurochem. 56:560 (1991); Oligodeoxynucleotides As Antisense Inhibitors Of Gene Expression, CRC Press, Boca Raton, FL (1988), for a description of these molecules).

In a particular aspect the invention provides the use of the polypeptide, polynucleotide or antagonist of the invention to interfere with the initial physical interaction between a pathogen and mammalian host responsible for sequelae of infection.

The invention further includes molecules which block the function of the polypeptides according to the invention or a polynucleotide encoding the same, identifiable by any of the above described methods.

An antagonist of the invention may be provided in pharmaceutical compositions which may include a carrier. They may be provided in unit dosage form. Such agents and pharmaceutical compositions are within the scope of the present invention. In order to prepare such pharmaceutical compositions the inhibitors will normally be provided in substantially pure form. They can then be combined with a carrier under sterile conditions.

The present invention also provides a method of treatment which comprises administering to a patient an effective amount of an antagonist of the expression or function of a polypeptide as defined herein.

The present invention further provides the use of an antagonist of a polypeptide as defined herein or a polynucleotide encoding the same for the manufacture of a medicament for the treatment of a bacterial infection.

Figures

Figure 1 shows the multiple sequence alignment and BLAST based identification of the ygjD family members.

Figures 2a-d show position-dependant scoring matrices for profile-based identification of ygjD family members.

Figure 3 shows the PROSITE patterns of ygjD family members based on the motifs generated from the position dependent scoring matrices.

Figure 4 shows the outline cloning strategy for a gene disruption plasmid. The black box represents the adapter sequence.

Figure 5 shows Growth dependence on arabinose of a conditional mutant in the *E coli* gene ygjD. An E. *coli* MG1655 derivative in which the chromosomal areBAD genes have been replaced with ygjD and the native ygjD gene has been deleted is shown on the upper half of each plate and a wild-type control is shown on the lower half of each plate.

Figure 6 is a diagram of the vector used to create conditional mutants in *B. subtilis.*

Figure 7 shows growth dependence on xylose of a conditional mutant in the *B. subtilis* ygjD orthologue yidE.

Figure 8 shows over-expression of the ygjD protein.
SDS-PAGE of E. *coli* MG1655/pASK-ygjD (Lanes 1, 3, 5) and MG1655/pASK75 (Lanes 2 and 4) whole-cell extracts. M-molecular weight standard. Lane 1 : uninduced. Lanes 2 and 3: 1 hour induction. Lanes 4 and 5: 3 hours induction.

Examples

Example 1. Identification of conserved bacterial open reading frames.

The predicted open reading frames obtained from the complete *E. coli* genomic sequence (Science (1997) 277: 1453-1474) were compared in a serial manner to the predicted open reading frames of the *H. influenzae* (Science (1995) 270:397-403), *M. genatilum* (Science (1995) 270:397-403), *Synechocystis* (Nuc. Acids Res. (1998) 26: 63-67) and *B. subtilis* (Nature (1997) 390:249-256) complete genome sequences

using the BLAST algorithm (J. Mol. Biol. (1990) 215:403-10). All matches with BLAST Score of greater than 75 were then analysed in a pair-wise fashion using the SIM algorithm (Advances in Applied Mathematics (1991) 12:337-357). The SIM score was then divided by a "selfSIM" score, a value obtained when the query protein is compared to itself using SIM algorithm with the PAM200 matix, to yield a similarity value of between 1.0 and 0. Proteins for which this similarity value was greater than 0.2 when the *E. coli* protein was compared to either the *B. subtilis* or *M. genatilum* genome where then compiled into a list and manually screened to identify proteins of unknown function. Those open reading frames which also had high similarity values in other bacteria were then considered as candidate genes and targets for gene disruption.

Example 2. Demonstration of essentiality of ygjD **genes in *E. coli.***

2A - In-frame deletion of selected genes in *E. coli*.

A disruption plasmid was constructed using DNA containing an in-frame deletion of the gene of interest plus ~900 base pairs of 5' and 3' flanking DNA for homologous recombination. The plasmid was cloned into the gene-replacement vector pKO3 as follows: Two separate PCR reactions were used to amplify fragments of approximately 900 base pairs of 5' and 3' sequence flanking the gene of interest. Chromosomal DNA from *E .coli* strain MG1655 was used as the template. Primers 2 and 3 carry a 5' extension of a 33 bp adapter sequence

adaptor sequence forward direction 5'-gttataaatttggagtgtgaaggttattgcgtg;
adaptor sequence reverse direction 5'-cacgcaataaccttcacactccaaatttataac.

Subsequently, the 2 PCR products were purified using High Pure™PCR Product Purification Kit (Boehringer Mannheim Inc., Mannheim, GE). Using the adapter sequence, the 2 PCR products are assembled in a second PCR reaction to give a single product . Following restriction enzyme digestion, preparative agarose gel electrophoresis and purification using Jetsorb™Gel Extraction Kit (Genomed Inc.) the final product was cloned into pKO3 using standard techniques. This clone is referred to as the disruption plasmid. All PCR reactions described in this section were performed with PWO™ DNA Polymerase (Boehringer Mannheim Inc.,

Mannheim, GE). In the final product the gene of interest was deleted from the start to the stop codon and replaced by the 33 bp adapter sequence [e.g. 5'-ATGgttataaatttggagtgtgaaggttattgcgtgTAA-3']. As a consequence the reading frame is maintained.

2B - Construction of an in-frame deletion mutant of *Escherichia coli*

The disruption vector pKO3 (A.J.Link et al., J. Bacteriol. 179:6228-6237,1997) is a derivative of pMAK700 (C.A.Hamilton et al., J. Bacteriol. 171:4617-4622). It features the *repA* (Ts) replication origin derived from pSC101 [permissive at 30°C but inactive at 42 to 44°C], the *cat* gene encoding chloramphenicol resistance and the sacB gene for counter selection against vector sequences in the presence of 5% sucrose.

The disruption plasmid described above was transformed into MG1655. Subsequently, chromosomal integrates (cointegrates produced by a single homologous recombination event) of the plasmid were isolated by selecting clones on chloramphenicol at 44°C. Following 2-times purification under the same conditions, the cointegrates are grown at 30°C in the presence of 5% sucrose to force resolution of the cointegrate and elimination of the plasmid from the cell. At this step, a preliminary assignment if a given gene is essential or non-essential for growth of *E. coli* in complex media was made. The genotype of the chloramphenicol-sensitive clones obtained following cointegration and resolution of the disruption plasmid was determined by colony-PCR using primers c1 and c2 (see Fig.4). In the case of a non-essential gene, the second recombination event can result in either a wild-type or a mutant genotype. The testing of 20 independent clones, showed routinely that a ~1:1 distribution of wild-type versus mutant genotype in case of a non-essential gene. Recovery of only wild-type genotype in 50 independent clones was considered as preliminary evidence for a gene's essentiality.

2C - Construction of a conditional mutant and final proof that a given gene is essential for growth of *E. coli*

A vector, pRDC15 was designed, which allows a copy of a putative essential gene to be placed in ectopic position on the chromosome under the control of a tightly

regulated promoter. The plasmid is a derivative of pKO3. In addition to the attributes of pKO3, pRDC15 carries a DNA fragment consisting of the *araC* gene, the arabinose promoter, a cloning site [*Bam*HI-*Nhe*I-*Sfi*I-*Xho*I-*Sph*I-*Sfi*I] and the *polB* gene. The wild-type copy of a putative essential gene was amplified by PCR and cloned into the vector pRDC15 using restriction sites *Nhe*I and *Xho*I. The resulting construct was used for gene replacement in a manner identical to the disruption plasmids described above. In this case the *araC* and *polB* genes of pRDC15 represent the homologous DNA for recombination at the *araCBADpolB* locus of the *E. coli* chromosome. Following cointegration and resolution, the *araBAD* genes in the *E. coli* chromosome are replaced by the wild-type copy of the gene of interest, which is now under the control of the arabinose promoter. This merodiploid strain is then used to construct an in frame deletion of the wild-type target gene using the disruption plasmid described above in the presence of 0.2% arabinose. In this case, the deletion mutant can be obtained since a wild-type copy is expressed in trans from the arabinose locus. The resulting strain is a conditional mutant as expression of the target gene is now dependent on the presence of arabinose. The inability of such a strain to grow in the absence of arabinose is a final proof that a given gene is essential for growth of *E. coli*. Figure 5 shows that the gene ygjD is essential in *E. coli*.

Example 3 yidE **is an essential gene in *Bacillus subtilis*.**

3A - Construction of a *B. subtilis* integrative plasmid for xylose controlled gene expression.

An integrative plasmid allowing the expression of genes under the control of a xylose inducible promoter was constructed as follows: A DNA fragment carrying the repressor gene *xylR* and the *xylA* promoter was PCR amplified from *B. subtilis* genomic DNA with the following primers:

pxyl-4: 5'-atcgctcgagAGATGCACCTTCTATACCCG-3'
pxyl-7: 5'-atcgaagcttAGCGATCCTACACAATCATG-3'

The primers were designed such that they introduced a unique *Eco*RI site at the 5' end of the PCR product and a unique *Bam*HI site at the 3' end of the product. The

PCR fragment was then cloned as an *Eco*RI-*BamHI* fragment into the *B. subtilis* integrative vector pDG648 to yield pRDC9 (Figure 6).

3B - Construction of the disruption plasmid.

A DNA fragment containing approximately 100 bp sequence from the 5' region of yidE was amplified by PCR from *B. subtilis* genomic DNA. The PCR primers were designed such that the resulting PCR product contains unique restrictions site at both the 5' and 3'ends of the PCR product. Subsequently, the PCR product was cloned into pRDC9.

3C - Construction of a conditional mutant.

The disruption plasmid was inserted into *B. subtilis* strain JH642. Chromosomal integration of the plasmid via single-reciprocal Campbell-like recombination at the yidE locus into the chromosome was driven by selection on LB plates containing erythromycin (1 μg/ml), lincomycin (25 μg/ml) and 10 mM xylose. The resulting strain is a conditional mutant in which expression of yidE is dependent on the presence of xylose into the growth medium.

3D - Confirmation that yidE is an essential gene.

Confirmation of that yidE is essential for growth was obtained by streaking the yidE conditional mutant LB plates plates containing erythromycin (1 μg/ml), lincomycin (25 μg/ml) with or without 10 mM xylose. The strain formed single colonies only on xylose containing plates thereby indicating that expression of yidE is indispensable for growth (Figure 7).

Example 4 - Characterisation of the ygjD **polypeptide family**

4A - Repetitive BLAST searches

Repetitive BLAST searches (Altschul, S.F., Gish, W., Miller, W., Myers E.W., and. Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10) in which each of the of the ygjD protein family members described below were used

in succession as query sequences to identify other members of the ygjD family as proteins which yield high-scoring segment pairs (HSP) scores of greater than 100 in comparison to at least one member of the ygjD polypeptide sequnces shown in figure 1 when a BLOSUM62 scoring matrix is used.

Sources for each of the sequences set out in Figure 1 are given below:

H. influenzae - GCP, Swissprot accession number P43764
P. haemolytica - GCP, Swissprot accession number P36175
E. coli - ygjD, Swissprot accession number P05852
M. leprae - Y246, Swissprot accession number P37969
M. tuberculosis - Y09A, Swissprot accession number Q50709
S. epidermidis - GlaxoWellcome *S. epidermidis* genomic sequencing project ORF Z0254002
B. subtilis - yidE, Swissprot/trEMBL accession number O05518
S. pyogenes - Contig229 from *S. pyogenes* genome sequencing project, B.A. Roe, S. Clifton, Mike McShan and Joseph Ferretti (http://www.genome.ou.edu/strep.html), August 25, 1997 data release
S. pneumoniae - GlaxoWellcome *S. pneumoniae* genomic sequencing project contig SP09_0003
Synechocystis - Y807, Swissprot accession number P74034
B. burgdorferi - EMBL accession number G2688702
T. pallidium - contig 6278 from the *T. pallidium* genome sequencing project at http://www.ncbi.nlm.nih.gov/BLAST/tigr_db.html
M. genitalium - GCP, Swissprot accession number P47292
M. pneumoniae - GCP, Swissprot accession number P75055
A. thaliana - F4L23.22, Swissprot/trEMBL accession number O22145
H. pylori - GCP, Swissprot accession number P55996

4B - Profile based searches

Multiple sequence alignments of the ygjD family members have been used to identify short patterns of amino acid sequences, which are common to all of the family members. Four motifs have been identified in the ygjD gene family using the

motif discovery tool, MEME (Bailey, T. L. and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994). Each of the four motifs are shown as they exist in each of the family members and are explicitly described as position-dependent scoring matrices, or profiles. Together these profiles can be used by the motif alignment and search tool, MAST, described in the same reference, to search databases for ygjD family members, which are positively identified when p-values of less than 1×10^{-50} are obtained. Where p-values are based on a random sequence model that assumes each position in a random sequence is generated according to the average letter frequencies of all sequences in the peptide non-redundant database (ftp://ncbi.nlm.nih.gov/blast/db/) on September 22, 1996.

Tables 1 to 4 show the position dependent scoring used to define the ygjD family. Values in the position-dependent scoring matrix are calculated by taking the log (base 2) of the ratio p/f at each position in the motif where p is the probability of a particular letter at that position in the motif, and f is the average frequency of that letter in the training set. Columns correspond to 1 letter amino acid codes and rows correspond to the position in the motif.

Table 1 - Position-dependent scoring matrix.
Values are the position-dependent scoring matrix are calculated by taking the log (base 2) of the ratio p/f at each position in the motif where p is the probability of a particular letter at that position in the motif, and f is the average frequency of that letter in the training set. Columns correspond to 1 letter amino acid codes and rows correspond to the position in the motif.

log-odds matrix: alength= 20 w= 53 n= 4663 bayes= 8.18208

	A	C	D	E	F	G	H	I	K	L
1	-2.277	-2.125	-3.382	0.004	3.402	-3.723	-0.659	-1.993	-3.110	-0.176
2	-2.790	2.439	-4.259	-4.185	-4.839	-4.269	-3.936	-4.722	-3.978	-4.550
3	1.160	-2.245	-4.754	-4.142	2.973	-4.026	-2.836	-1.096	-3.812	1.365
4	0.642	-1.760	-4.278	-3.633	-1.708	-3.518	-2.605	1.188	-3.305	1.586
5	3.059	0.873	-3.708	-3.369	-3.069	0.388	-3.276	-2.698	-3.370	-0.602
6	-5.188	-4.668	-6.364	-6.163	-3.267	-5.996	-5.157	-2.799	-5.924	3.389
7	-4.070	-3.547	-6.721	-6.271	-2.812	-6.438	-5.688	0.547	-6.117	2.679
8	-0.452	-1.804	-3.825	-3.551	-2.668	-3.980	-2.950	1.142	-3.556	-0.952
9	-1.550	1.609	-4.753	-5.200	-4.945	0.701	-4.509	-5.021	-4.922	-5.171
10	-3.360	-4.471	-3.904	-4.672	-5.558	3.790	-4.304	-5.722	-4.493	-6.020
11	-2.691	-3.918	-3.223	-3.992	-4.964	3.647	1.616	-5.091	-3.819	-5.476
12	-4.076	-3.825	-3.648	-4.266	-3.422	-4.479	5.201	-5.151	-4.574	-4.412
13	-0.646	-2.490	-3.419	-3.955	-3.899	-4.111	1.688	-2.882	-3.229	-4.017
14	0.626	-3.788	-1.207	1.999	-3.851	-3.219	-2.055	-3.353	-1.480	1.675
15	-4.068	-3.565	-6.355	-5.758	-2.430	-6.159	-5.036	1.995	-5.569	2.848
16	-0.228	1.200	-3.892	-3.403	-1.867	-3.499	0.827	1.004	-3.198	-0.397
17	-1.492	-2.671	-1.724	0.151	-1.514	-2.679	1.221	-2.316	0.981	0.619
18	0.278	-1.661	-3.702	-3.277	-2.100	-3.580	-2.555	0.398	-3.059	-1.626
19	-1.479	-3.129	1.022	-0.928	-3.126	-2.526	2.098	-2.989	2.335	-2.856
20	-1.894	-3.496	1.153	1.442	-3.579	1.514	-1.434	-3.512	0.298	0.150
21	-0.132	1.467	-3.955	-3.378	1.924	0.465	2.079	0.461	-3.090	0.468

22	-0.420	-3.097	0.637	0.482	-0.104	2.935	-1.867	-3.009	-1.591	-1.074
23	-2.008	-3.702	2.564	0.441	-3.781	-2.686	-1.600	-3.669	0.267	-3.543
24	-2.922	-2.766	-4.685	-4.318	1.595	-4.565	-2.016	1.575	-4.146	0.530
25	-1.603	-3.194	-1.606	2.779	-3.206	-2.719	1.246	-0.139	1.112	-2.868
26	-0.119	-1.766	-3.911	0.669	-1.724	-3.441	-2.495	1.994	-3.040	1.281
27	-4.069	-3.541	-6.837	-6.474	-2.965	-6.536	-5.986	2.675	-6.336	2.353
28	0.639	-3.282	-3.220	-3.936	-4.854	3.571	-3.653	-4.877	-3.783	-5.259
29	-0.126	-3.711	-1.318	2.957	-4.087	-3.083	-2.038	-3.766	-1.480	-3.677
30	-2.465	-2.499	-3.869	-4.301	-4.038	-4.046	-3.475	-3.038	-3.497	-4.162
31	-1.923	-1.853	-4.318	-3.674	-1.817	-3.611	-2.693	2.001	-3.257	0.962
32	-4.178	-4.465	4.179	-2.009	-4.984	-4.496	-3.386	-4.990	-4.723	-5.260
33	-4.178	-4.465	4.179	-2.009	-4.984	-4.496	-3.386	-4.990	-4.723	-5.260
34	3.539	-1.630	-4.458	-4.234	-3.994	-2.384	-4.048	-3.741	-4.276	-3.889
35	2.364	1.487	-4.265	-3.666	0.482	-3.301	-2.731	1.578	-3.376	-1.465
36	-3.360	-4.471	-3.904	-4.672	-5.558	3.790	-4.304	-5.722	-4.493	-6.020
37	-4.990	-5.543	-2.174	3.953	-6.011	-5.282	-4.480	-5.343	-5.611	-6.223
38	3.375	-1.078	-3.842	-3.525	-3.291	-1.964	-3.448	-2.967	-3.542	-3.133
39	-5.599	-4.474	-6.036	-6.142	3.562	-5.995	-2.205	-4.624	-5.622	-4.083
40	-4.178	-4.465	4.179	-2.009	-4.984	-4.496	-3.386	-4.990	-4.723	-5.260
41	-3.922	-3.914	-4.975	-4.639	-5.549	-4.909	-3.866	-4.342	3.994	-5.076
42	-3.024	-2.944	-5.684	-5.451	-3.763	-5.703	-5.208	0.648	-5.446	-2.598
43	3.044	-2.339	-5.290	-5.379	-5.125	2.044	-4.872	-5.052	-5.518	-5.185
44	-0.334	-4.447	-4.487	-3.045	-5.439	-4.316	-2.191	-4.393	2.627	-4.062
45	0.619	-1.692	-4.165	-3.514	0.538	-3.424	1.138	-0.720	-3.188	1.841
46	-4.570	-4.048	-7.081	-6.391	-2.019	-6.503	-5.054	0.354	-6.140	2.449
47	-2.436	-3.689	0.903	-3.248	-4.738	3.510	-2.944	-4.849	-3.258	-5.223
48	-4.278	-3.779	-6.179	-5.450	1.876	-6.012	-4.504	-1.494	-5.229	2.998
49	-2.805	-3.827	2.487	-2.581	-5.231	2.284	-2.410	-5.697	-3.113	-5.633
50	-4.602	-3.990	-5.305	-5.264	1.052	-5.486	-1.793	-4.051	-4.975	-3.586
51	-2.431	-3.879	-3.759	-3.650	-4.361	-4.009	0.988	-4.244	-3.439	-4.062
52	1.632	-2.934	-3.618	-4.253	-5.002	3.277	-3.921	-5.007	-4.003	-5.300

53	-2.234	-3.508	-2.862	-3.609	-4.666	3.685	-3.356	-4.688	-3.438	-5.123
	M	N	P	Q	R	S	T	V	W	Y
1	0.629	-3.048	-0.650	-3.014	-3.013	-2.583	-2.753	-2.018	-1.203	2.865
2	-4.331	-4.379	4.056	-3.764	-4.139	-3.376	-3.674	-4.248	-5.181	-5.127
3	-0.923	-3.679	-4.137	-3.383	-3.631	-3.166	-2.346	0.993	-2.479	0.868
4	2.225	-3.208	-3.667	-2.949	-3.156	-0.332	-1.817	1.824	-2.452	-2.147
5	-2.055	-3.375	-4.260	-3.313	-3.308	0.208	-2.102	-0.382	-3.297	-3.545
6	-1.990	-6.098	-5.693	-5.136	-5.439	-5.939	-5.128	-3.485	-4.622	-4.737
7	-1.594	-6.068	-6.034	-5.455	-5.959	-5.873	-4.038	2.211	-4.645	-4.563
8	-1.660	-3.650	-3.595	-3.492	-3.252	-0.361	-1.843	3.314	-3.575	-3.550
9	-4.168	-3.550	-4.137	-4.453	-4.613	3.358	-1.565	-3.834	-5.042	-5.032
10	-4.983	-3.645	-5.067	-4.864	-4.390	-3.867	-4.815	-5.224	-4.787	-5.149
11	-4.332	-2.929	-4.555	-4.226	-3.756	-3.192	-4.183	-4.589	-4.240	-4.506
12	-3.651	-2.191	-4.623	-1.877	-3.012	-3.587	0.698	-4.632	-3.513	-1.844
13	-2.380	0.008	-4.033	-2.964	-3.315	0.163	3.521	-2.544	-3.940	-4.150
14	-2.565	-2.067	-3.245	1.970	-2.141	-0.231	-2.304	-2.935	-4.012	-3.319
15	-1.157	-5.716	-5.546	-4.768	-5.233	-5.505	-4.013	0.211	-4.146	-4.185
16	-0.880	-3.174	-3.459	0.015	-2.995	-2.667	-1.665	2.837	-2.610	1.259
17	-1.619	-1.448	-2.741	-0.927	0.275	-0.159	-1.591	1.111	-2.544	2.600
18	1.615	-3.203	-3.420	-0.209	0.368	-2.774	-1.667	3.085	-2.861	-2.673
19	-1.954	-1.164	0.007	-0.736	0.359	1.081	0.091	-2.558	-3.166	-2.443
20	-2.503	1.337	-3.003	-1.165	-1.760	0.659	-1.903	-3.070	-3.662	-2.846
21	-0.667	-3.008	-0.049	-2.775	-2.952	-2.459	-1.737	0.993	-2.115	1.197
22	-2.311	-1.698	-3.220	-1.619	-2.005	-1.952	-2.189	-2.696	-3.325	0.214
23	-2.665	0.564	0.984	2.338	-1.926	-0.114	-2.061	-3.205	-3.858	-3.034
24	-1.232	-4.028	-4.464	-3.783	-3.955	-3.612	-3.076	0.884	-2.293	3.588
25	-1.992	-1.415	-2.798	-0.801	0.407	-1.612	0.072	-0.233	-3.229	-2.553
26	-0.712	-3.043	-3.588	1.273	-2.994	-2.549	-1.777	1.478	-2.444	-2.126
27	-1.770	-6.187	-6.234	-5.773	-6.300	-6.007	-4.041	1.141	-4.856	-4.709
28	-4.099	-2.880	-4.283	-4.085	-3.685	-2.699	-3.592	-4.139	-4.187	-4.497

29	-2.797	-1.935	-3.271	-1.279	0.212	0.614	1.403	-3.213	-4.132	-3.410
30	-2.533	-2.312	-4.117	-3.189	-3.520	1.400	3.596	-2.656	-4.075	-4.335
31	-0.806	-3.277	-3.764	-2.998	1.921	0.528	-1.910	1.508	-2.571	-2.260
32	-4.644	-1.840	-5.295	-4.287	-4.666	-4.198	-4.647	-4.778	-4.692	-4.552
33	-4.644	-1.840	-5.295	-4.287	-4.666	-4.198	-4.647	-4.778	-4.692	-4.552
34	-3.013	-3.921	-4.329	-3.997	-4.100	-0.130	-2.523	-2.253	-4.158	-4.433
35	-0.874	-3.283	-0.047	-3.053	-3.241	-2.422	-1.891	1.451	-2.596	-2.309
36	-4.983	-3.645	-5.067	-4.864	-4.390	-3.867	-4.815	-5.224	-4.787	-5.149
37	-5.471	-4.281	-6.363	-3.761	-5.565	-5.496	-5.414	-5.556	-5.666	-5.816
38	-2.280	-3.493	-4.295	-3.461	-3.462	-0.236	-2.193	0.269	-3.484	-3.771
39	-3.974	-4.668	-5.909	-4.677	-5.028	-5.239	-5.330	-4.709	-1.526	3.899
40	-4.644	-1.840	-5.295	-4.287	-4.666	-4.198	-4.647	-4.778	-4.692	-4.552
41	-4.147	-3.854	-4.789	-4.069	-0.855	-4.705	-4.183	-4.936	-4.366	-4.942
42	-2.532	-5.379	-5.446	-5.487	-5.390	-5.081	1.797	3.355	-5.511	-4.990
43	-4.227	-4.342	-4.293	-4.669	-5.039	-0.030	-2.822	-3.476	-5.220	-5.455
44	-3.351	-2.960	-4.583	-1.535	3.296	-3.588	-3.355	-4.107	-4.223	-4.017
45	1.393	-3.104	-3.561	-2.826	-3.036	-2.545	0.869	1.023	1.846	-2.050
46	3.602	-6.224	-5.887	-4.920	-5.605	-5.973	-4.523	-2.478	-3.734	1.743
47	-4.095	-0.153	-4.167	-3.517	-3.389	-2.604	-3.566	-4.359	-4.079	-4.148
48	1.385	-5.527	-5.192	-4.245	-4.737	-5.316	-4.189	-2.312	-3.574	-3.686
49	-4.986	-0.991	1.713	-2.937	-3.821	-2.216	0.966	-4.894	-5.391	-4.270
50	1.282	-4.334	-5.332	-4.332	-4.554	-4.557	-4.747	-4.146	-1.445	4.633
51	-3.853	-3.899	4.113	-3.248	-3.639	-2.995	-3.261	-3.821	-4.815	-4.635
52	-4.204	-3.174	-4.263	-4.166	-0.217	-2.485	-3.258	-3.983	-4.514	-4.772
53	-3.919	-2.560	-4.206	-3.861	-3.374	-1.101	-3.767	-4.162	-3.908	-4.224

Table 2 - Position-dependent scoring matrix.
Values are the position-dependent scoring matrix are calculated by taking the log (base 2) of the ratio p/f at each position in the motif where p is the probability of a particular letter at that position in the motif, and f is the average frequency of that letter in the training set. Columns correspond to 1 letter amino acid codes and rows correspond to the position in the motif.
log-odds matrix: alength= 20 w= 47 n= 4759 bayes= 8.21158

	A	C	D	E	F	G	H	I	K	L
1	-2.804	-5.908	3.400	1.498	-5.852	-3.954	-3.127	-4.840	0.889	-4.770
2	-3.256	-3.087	-4.606	-4.579	-2.676	-4.939	-4.330	3.476	-4.187	0.596
3	-2.817	-4.360	3.273	0.404	-4.744	-2.841	-2.130	-4.727	0.083	-4.632
4	2.925	-1.132	-3.704	-3.261	-2.352	0.596	1.490	-0.226	-3.140	-0.812
5	-3.726	-3.281	-5.911	-5.834	-3.721	-6.070	-5.978	3.409	-5.596	-2.146
6	3.728	-3.537	-5.976	-6.166	-5.773	-4.074	-5.500	-5.746	-6.165	-5.826
7	0.656	-2.774	-5.641	-5.161	-2.935	-5.014	-4.307	0.632	-4.909	-2.240
8	0.709	-2.602	-4.709	-5.032	-4.743	-3.400	-4.280	-3.979	-4.499	-4.844
9	1.085	1.320	-1.681	0.190	-2.305	-2.613	1.207	1.323	-0.930	-2.192
10	-0.215	-3.540	-2.599	-3.338	-4.606	3.459	-3.101	-4.621	-3.082	-5.002
11	-4.452	-5.372	-5.480	-5.619	-6.031	-5.391	-5.110	-6.194	-5.399	-5.906
12	-3.360	-4.471	-3.904	-4.672	-5.558	3.790	-4.304	-5.722	-4.493	-6.020
13	-5.188	-4.668	-6.364	-6.163	-3.267	-5.996	-5.157	-2.799	-5.924	3.389
14	1.373	-1.780	-4.026	-3.545	-2.128	-3.390	-2.741	1.026	-3.347	-1.702
15	-2.230	-3.488	-2.850	-3.589	-4.537	3.644	-3.339	-4.285	-3.413	-1.815
16	3.109	2.945	-4.621	-4.407	-4.147	-2.441	-4.179	-3.906	-4.456	-4.055
17	-5.188	-4.668	-6.364	-6.163	-3.267	-5.996	-5.157	-2.799	-5.924	3.389
18	-3.205	-2.866	-4.763	-4.027	-1.693	-4.738	1.182	0.416	-3.774	2.964
19	-1.262	-1.800	-3.728	-1.297	-2.944	-4.088	-3.006	0.715	-3.589	-2.215
20	-3.360	-4.471	-3.904	-4.672	-5.558	3.790	-4.304	-5.722	-4.493	-6.020

21	1.544	-1.993	-4.568	-3.949	-2.018	-3.794	-2.963	0.556	-3.634	1.594
22	0.756	-2.511	-2.996	-2.884	-3.628	-3.727	-2.638	-2.774	-2.413	-3.599
23	2.282	-1.723	-4.187	-3.553	2.294	-3.418	-2.578	1.164	-0.154	-1.303
24	3.557	-1.839	-4.645	-4.459	-4.215	-0.278	-4.235	-3.980	-4.503	-4.122
25	-3.706	-4.512	-4.596	-3.163	-5.505	-4.411	-2.304	-4.457	3.328	-4.160
26	2.643	-2.068	-4.329	-4.220	-4.485	-0.179	-4.032	-4.295	-4.147	-4.424
27	-3.998	-3.534	-5.654	-5.019	1.302	-5.525	-3.462	-1.608	-4.761	2.902
28	3.040	-2.449	-5.461	-5.558	-5.277	-2.918	-5.000	-5.235	-5.703	-5.371
29	0.458	-1.728	-3.583	-3.160	2.582	-0.585	-1.082	0.258	-2.968	-0.313
30	3.174	-0.939	-3.610	-3.267	-3.003	-1.867	-3.186	-2.617	-3.258	-0.654
31	-4.227	-3.721	-5.522	-5.337	0.971	-5.315	2.358	-3.281	-4.932	1.321
32	-2.977	-4.173	2.164	-2.160	-4.673	1.280	2.297	-4.993	0.128	-4.855
33	-3.251	-2.953	-5.793	-5.220	-2.829	-5.262	-4.450	0.599	2.206	1.676
34	-4.452	-5.372	-5.480	-5.619	-6.031	-5.391	-5.110	-6.194	-5.399	-5.906
35	1.192	-3.439	-6.336	-5.701	2.483	-5.606	-4.515	1.702	-5.412	2.020
36	-3.319	-3.105	-4.940	-4.816	-2.552	-5.075	-4.354	3.234	-4.448	1.096
37	1.344	1.635	-4.853	-4.786	-4.756	1.990	-4.366	-4.465	-4.713	-0.685
38	-1.475	-1.962	-3.950	-1.034	-3.068	-4.292	-3.242	1.175	-3.804	-2.280
39	-3.552	-4.473	1.336	-2.576	-5.184	-2.594	4.096	-6.021	-2.946	-5.853
40	-6.822	-5.947	-6.129	-6.963	-5.889	-6.051	5.460	-7.656	-7.005	-6.859
41	-4.861	-4.246	-7.656	-6.983	-2.099	-6.979	-5.540	1.707	-6.757	2.257
42	1.422	-2.965	-1.402	1.974	-2.907	0.471	1.225	-2.723	-0.657	-0.702
43	1.353	-3.082	-3.750	-4.413	-5.102	3.453	-4.071	-5.131	-4.297	-5.438
44	-6.822	-5.947	-6.129	-6.963	-5.889	-6.051	5.460	-7.656	-7.005	-6.859
45	-0.371	-2.912	-5.520	-4.889	0.514	-4.980	-3.964	2.267	-4.615	2.423
46	0.422	1.289	-3.602	-3.240	1.603	-3.547	-0.935	-1.251	-3.089	1.072
47	3.317	-1.191	-3.972	-3.668	-3.433	-2.050	-3.573	-3.124	-3.689	-3.284

	M	N	P	Q	R	S	T	V	W	Y
1	-3.976	-2.587	-3.891	1.437	-3.426	-3.323	-3.449	-4.134	-5.746	-4.883
2	-1.002	-4.155	0.308	-4.219	-4.442	-4.042	-3.035	0.781	-3.876	-3.426
3	-3.854	-1.045	-3.761	-1.958	-2.944	1.147	0.855	-4.229	-4.820	-3.828
4	-1.346	-3.168	-3.949	-2.999	-3.091	-1.305	-1.918	-0.057	-2.841	-2.778
5	-2.189	-5.455	-5.997	-5.742	-5.963	-5.367	-3.639	2.442	-5.526	-4.763
6	-5.120	-5.548	-5.455	-5.683	-5.783	-3.470	-4.248	-4.400	-5.653	-5.987
7	-1.960	-4.748	-5.145	-4.657	-4.869	-4.236	-3.052	2.866	-4.060	3.039
8	-3.458	-3.301	-4.316	-4.016	-4.364	-1.556	3.764	-3.271	-4.798	-5.031
9	-1.415	0.370	-2.699	0.646	-1.398	0.580	-1.468	-0.072	-2.768	1.538
10	-3.840	0.314	-4.126	-0.572	-0.803	-2.667	-3.613	-4.126	-3.888	-4.103
11	-5.812	-5.700	4.270	-5.176	-5.434	-4.964	-5.188	-5.779	-5.952	-6.170
12	-4.983	-3.645	-5.067	-4.864	-4.390	-3.867	-4.815	-5.224	-4.787	-5.149
13	-1.990	-6.098	-5.693	-5.136	-5.439	-5.939	-5.128	-3.485	-4.622	-4.737
14	-1.134	-3.295	-0.181	-3.075	-3.183	0.339	-0.055	2.639	-2.853	-2.608
15	-3.763	-2.550	-4.198	-3.841	-3.351	-2.764	-3.750	-1.247	-3.871	-4.168
16	-3.169	-4.007	-4.289	-4.096	-4.237	0.621	-0.112	-2.413	-4.308	-4.582
17	-1.990	-6.098	-5.693	-5.136	-5.439	-5.939	-5.128	-3.485	-4.622	-4.737
18	0.948	-4.143	-3.985	-3.081	-0.824	-3.865	-3.094	-1.618	-3.048	-3.055
19	-1.937	-3.740	-3.562	-3.607	-3.249	-3.428	-1.832	3.556	-3.936	-4.101
20	-4.983	-3.645	-5.067	-4.864	-4.390	-3.867	-4.815	-5.224	-4.787	-5.149
21	-0.995	-3.525	-3.971	-3.291	-3.501	-0.334	-2.088	2.151	-2.810	-2.491
22	0.837	1.390	-3.692	-2.307	-0.229	0.125	3.088	-2.440	-3.717	-3.639
23	1.388	-3.163	-3.653	-2.902	-3.109	-2.537	-1.802	0.422	-2.449	-2.145
24	-3.251	-4.096	-4.450	-4.193	-4.306	-1.756	-2.703	-2.493	-4.359	-4.639
25	-3.444	-3.059	-4.664	-1.651	2.717	-3.712	-3.457	-4.206	-4.294	-4.104
26	-3.520	-3.695	-4.179	-3.853	0.033	1.886	0.894	-2.885	-4.622	-4.720
27	-0.798	-4.859	-4.865	-3.916	-4.309	-4.738	-3.889	-0.191	-2.788	1.535

28	-4.406	-4.462	-4.331	-4.777	-5.169	2.262	-2.893	-3.657	-5.356	-5.582
29	1.182	-2.895	-3.375	-2.734	-2.829	-2.378	-1.906	-0.965	3.201	2.408
30	-1.987	-0.754	-4.235	-3.216	-3.221	0.521	-2.067	-0.394	-3.237	-3.468
31	1.328	0.313	-5.306	-4.243	-4.539	-4.475	-4.149	-3.286	4.389	2.159
32	-4.117	2.364	-3.793	0.575	-2.979	-0.071	-2.662	-4.558	-4.813	-3.691
33	-1.696	-4.877	-5.283	-4.535	-4.569	-4.495	-3.229	2.240	-4.143	-3.827
34	-5.812	-5.700	4.270	-5.176	-5.434	-4.964	-5.188	-5.779	-5.952	-6.170
35	-0.982	-5.347	-5.434	-4.603	-5.099	-4.891	-3.735	-1.889	-3.593	-3.572
36	-1.088	-4.399	-5.045	-4.340	-4.605	-4.216	-3.153	1.110	-3.824	1.279
37	-3.923	-4.204	2.884	-4.140	-4.552	-2.252	-2.801	-3.471	-5.051	-5.082
38	-2.033	-3.942	-3.782	-3.824	-3.481	-3.634	-2.021	3.506	-4.132	-4.231
39	-5.373	3.138	-4.158	-2.803	-3.708	-2.210	-3.025	-5.584	-5.387	-4.005
40	-6.587	-5.318	-6.304	-5.037	-5.811	-6.446	-6.549	-7.282	-5.682	-4.761
41	3.893	-6.811	-6.258	-5.270	-6.121	-6.654	-4.822	-2.476	-3.918	-4.200
42	-1.778	-1.189	-2.577	-0.679	0.327	-1.368	-1.450	-2.332	-3.030	1.471
43	-4.340	-3.334	-4.381	-4.384	-4.137	-2.642	-3.431	-4.130	-4.582	-4.879
44	-6.587	-5.318	-6.304	-5.037	-5.811	-6.446	-6.549	-7.282	-5.682	-4.761
45	-0.960	-4.621	-4.806	-4.040	-4.373	-4.175	-3.135	0.230	-3.466	-3.343
46	1.981	-3.019	-3.496	-2.873	-2.942	-2.529	-2.186	-1.265	-1.557	3.258
47	-2.426	-3.585	-4.293	-3.572	-3.592	0.773	-2.260	-0.323	-3.621	-3.907

Table 3 - Position-dependent scoring matrix.
Values are the position-dependent scoring matrix are calculated by taking the log (base 2) of the ratio p/f at each position in the motif where p is the probability of a particular letter at that position in the motif, and f is the average frequency of that letter in the training set. Columns correspond to 1 letter amino acid codes and rows correspond to the position in the motif.
log-odds matrix: alength= 20 w= 11 n= 5335 bayes= 8.47031

	A	C	D	E	F	G	H	I	K	L
1	-3.740	-3.233	-6.342	-6.189	-3.848	-6.321	-6.476	3.182	-6.057	-2.419
2	-5.005	-4.499	-6.222	-5.972	-3.096	-5.877	-4.993	-2.619	-5.730	3.381
3	1.589	-2.653	-4.004	-4.589	-5.056	3.277	-4.202	-5.066	-4.534	-5.318
4	-3.206	-3.068	-4.358	-4.393	-0.366	-4.827	-4.258	3.706	-3.985	-0.826
5	-4.978	-5.535	-2.164	3.953	-6.002	-5.273	-4.471	-5.334	-5.596	-6.213
6	-2.469	-2.432	-3.771	-4.199	-3.930	-4.072	-3.369	-2.914	-3.378	-4.053
7	-1.749	-1.862	-3.302	-3.882	-3.501	-3.231	-3.084	-3.670	-3.033	-3.863
8	-3.840	5.729	-5.290	-5.017	-4.804	-5.450	-4.630	-4.265	-5.437	-4.880
9	-4.050	-4.350	4.170	-1.875	-4.865	-4.384	-3.261	-4.861	-4.598	-5.139
10	-2.788	-6.596	2.516	3.326	-6.256	-4.075	-3.342	-4.956	-2.424	-4.860
11	-2.411	1.173	-3.780	-4.192	-3.907	-4.029	-3.357	-2.888	-3.364	-4.026

	M	N	P	Q	R	S	T	V	W	Y
1	-2.450	-5.841	-6.218	-6.196	-6.446	-5.730	-3.701	2.787	-5.980	-5.095
2	-1.806	-5.926	-5.539	-4.947	-5.256	-5.757	-4.943	-3.297	-4.462	-4.566
3	-4.240	-3.461	-4.184	-4.354	-4.299	-0.205	-3.016	-3.822	-4.713	-4.982
4	-0.892	-3.962	-4.813	-4.102	-4.301	-3.905	-2.936	0.907	-3.838	-3.297
5	-5.460	-4.271	-6.352	-3.749	-5.554	-5.485	-5.403	-5.546	-5.658	-5.807
6	-2.413	-2.199	-4.045	-3.078	-3.407	0.737	3.724	-2.550	-3.967	-4.231

7	-2.823	-1.786	-3.328	-3.253	-2.948	3.352	1.205	-3.720	-3.648	-3.369
8	-3.926	-4.993	-5.489	-5.107	-4.869	-4.772	-4.042	-4.735	-5.452	-5.385
9	-4.516	-1.706	-5.192	-4.161	-4.546	-4.070	-4.523	-4.650	-4.580	-4.431
10	-4.065	-3.009	-3.867	-1.516	-3.722	-3.456	-3.528	-4.174	-6.178	-5.187
11	-2.385	-2.187	-4.017	-3.058	-3.388	0.211	3.721	-2.518	-3.946	-4.216

Table 4 - Position-dependent scoring matrix.

Values are the position-dependent scoring matrix are calculated by taking the log (base 2) of the ratio p/f at each position in the motif where p is the probability of a particular letter at that position in the motif, and f is the average frequency of that letter in the training set. Columns correspond to 1 letter amino acid codes and rows correspond to the position in the motif.

log-odds matrix: alength= 20 w= 21 n= 5175 bayes= 8.50403

	A	C	D	E	F	G	H	I	K	L
1	-1.732	-1.018	-3.179	-3.756	-3.356	-3.231	-2.950	-3.523	-2.879	-1.943
2	-2.797	-3.743	-1.585	-1.951	-4.572	-2.832	-1.387	-4.470	-1.691	-4.063
3	0.517	-2.065	-3.620	-0.405	-2.014	-3.675	-2.790	2.929	-3.035	-1.075
4	0.038	-3.075	1.643	1.641	-3.033	-2.454	-1.077	-2.883	1.589	-2.756
5	-1.412	-2.012	-2.141	1.396	0.526	-2.769	1.305	-1.215	-1.363	1.350
6	-3.365	-2.736	-2.476	-3.011	-2.202	-3.725	5.117	-3.927	-3.538	-2.603
7	2.305	-2.552	-2.134	-1.495	-3.268	-2.669	-1.612	-2.861	1.971	-2.902
8	-1.398	-3.040	1.622	-0.836	-3.045	-2.497	-1.049	-0.051	1.223	-2.740
9	-4.413	-3.824	-5.122	-5.086	3.278	-5.298	-1.625	-3.993	-4.799	-3.584
10	-2.181	-3.474	-2.732	-3.477	-4.479	3.641	-3.219	-4.570	-3.285	-5.008
11	-3.126	-4.260	-3.677	-4.442	-5.351	3.780	-4.090	-5.496	-4.263	-5.813
12	-2.203	-2.440	-4.747	-4.534	-3.407	-5.007	-4.145	1.270	-4.589	-2.377
13	-1.243	-2.158	-4.250	-4.053	-3.290	-4.260	-3.569	-0.386	-4.118	-2.461
14	-4.053	-5.066	-5.144	-5.237	-5.706	-5.103	-4.785	-5.811	-5.020	-5.545
15	-4.917	-5.513	-2.104	3.950	-5.972	-5.240	-4.424	-5.284	-4.915	-6.169
16	-3.160	-2.924	-5.026	-4.807	-2.464	-4.983	-4.254	2.866	-4.459	1.582
17	3.715	-3.158	-5.651	-5.761	-5.397	-3.740	-5.190	-5.307	-5.767	-5.415
18	0.604	-2.092	-3.660	-2.294	-4.106	-2.949	-3.578	-4.231	-3.678	-4.406
19	-3.499	-2.769	-4.342	-3.584	-4.600	-4.111	-1.868	-4.024	-1.290	-4.016
20	0.838	1.382	1.594	-0.694	-2.824	-2.341	2.709	-2.663	0.552	0.197
21	-6.430	-5.608	-5.724	-6.506	-5.443	-5.795	5.453	-7.231	-6.616	-6.437

	M	N	P	Q	R	S	T	V	W	Y
1	-2.678	-1.639	-3.219	-3.125	-2.798	3.291	1.233	-3.660	-3.507	-3.220
2	-3.099	0.646	-3.717	3.818	-1.135	1.547	-2.732	-4.126	-4.145	-3.715
3	-0.750	-3.135	-3.782	0.134	-3.139	-2.777	-2.007	1.868	-2.796	-2.429
4	-1.849	0.542	0.117	0.842	-1.179	-1.342	0.194	-2.435	-3.101	-2.365
5	1.375	-1.759	-2.873	-1.304	-1.741	1.163	0.227	-1.099	-2.491	-2.014
6	-2.278	-0.832	-3.519	0.326	-1.643	-2.479	-2.709	-3.440	-2.266	-0.478
7	-2.075	-1.793	-3.124	0.715	-1.176	-0.034	-1.870	0.910	-3.322	-2.812
8	-1.842	-1.174	0.122	1.032	2.334	0.024	-1.467	-2.451	-3.053	-2.356
9	-3.363	-4.157	-5.162	-4.168	-4.383	-4.369	1.035	-4.022	-1.284	3.728
10	-3.803	-2.433	-4.118	-3.734	-0.905	-2.696	-3.720	-4.081	-3.760	-0.432
11	-4.754	-3.412	-4.866	-4.643	-4.167	-3.635	-4.592	-4.997	-4.582	-4.937
12	-2.257	-4.653	-4.561	-4.630	-4.357	-4.380	-2.631	3.609	-4.847	-4.660
13	-2.244	-4.179	-4.022	-4.105	-3.803	-3.620	-2.267	3.721	-4.393	-4.425
14	-5.429	-5.351	4.261	-4.802	-5.085	-4.577	-4.811	-5.394	-5.702	-5.872
15	-5.405	-4.216	-6.320	-3.680	-5.501	-5.430	-5.349	-5.492	-5.635	-5.774
16	-4.681	-5.216	-5.210	-5.338	-5.433	-3.093	-3.911	-3.950	-5.334	-5.646
18	-3.387	-2.392	-3.656	-3.659	-3.577	3.366	-0.206	-3.678	-4.238	-4.029
19	-3.629	-3.218	-3.653	-2.313	4.121	-3.725	-3.774	-4.669	-2.934	-4.050
20	-1.668	-1.024	-2.423	0.843	-1.026	0.057	-1.302	-2.249	-2.912	-2.188
21	-6.056	-4.748	-6.017	-4.442	-5.317	-5.981	-6.113	-6.841	-5.301	-4.215

4C - PROSITE based searches

The conserved sequence elements identified with MEME can also be represented as PROSITE patterns using the conventions outlined in PROSITE: A dictionary of protein sites and patterns (http://www.expasy.ch/sprot/prosite.html) and Bairoch A., Bucher P., Hofmann K. The PROSITE datatase, its status in 1995. Nucleic Acids Res. 24:189-196(1995). YgjD family members are positively identified when exact matches to any one of the four prosite patterns pattern 1, pattern 2, pattern 3 or pattern 4 as set out in Figure 3 are found in the protein sequence. Alternatively, *ygjD* family members can be identified using PROSSITE pattern PS01016 found in the PROSITE database.

Example 5 - Over-expression of the *E. coli* ygjD polypeptide

The *E. coli* ygjD gene was amplified from *E. coli* chromosomal DNA in the presence of 1 µM each of the primers ask-eygjD5 [5'-gatctctagataaagcgaggtaaaacaagtc-3'] and ask-eygjD3 [5'-gatcctcgagtTTAcgcagccggtaactc-3'] and a nucleotide concentration of 250 µM using Pwo DNA Polymerase (Boehringer, Mannheim, Germany). 25 cycles of 30 sec at 94 °C/30 sec at 58 °C/1 min at 72 °C with a final 5 min extension at 72 °C were performed. The purified PCR product was cleaved with XbaI and XhoI and cloned into the expression vector pASK75 (Gene (1994) 151:131-135) cut with the same restriction endonucleases. The cloned ygjD gene was sequenced. The resulting plasmid pASK-ygjD was transformed into *E. coli* MG1655. Each 50 ml of LB medium containing 100 µg/ml carbenicillin was inoculated with 0.5 ml of a MG1655/pASK-ygjD or MG1655/pASK75 over-night culture and incubated at 30 °C. At an optical density of 0.65 at 600 nm, the cultures were induced with 200 ng/ml anhydrotetracycline. At the time of induction and after 1 and 3 hours post induction samples of 1 ml were withdrawn, the cells harvested by centrifugation and resuspended in 1x SDS-PAGE sample buffer (140 µl per 1 OD_{600} equivalent). The samples were boiled for 5 minutes and analyzed on a 4-20% SDS-PAGE gradient gel stained with Coomassie Brilliant Blue. Induction of a 36 kDa protein representing YgjD can be seen 1 and 3 hours following induction (Figure 8).

CLAIMS

1. An isolated polypeptide of the ygjD family as defined by:

i) an HSP score of greater than or equal to 100 when compared with one of the sequences of Figure 1 when the BLAST algorithm is used with a BLOSUM62 scoring matrix ; or

ii) containing a set of amino acid sequences which are positively identified when position dependent scoring matrices according to Tables 1-4 are used with MAST to yield a p-value of less than $1x10^{-50}$; or

iii) comprising at least one of the following amino acid sequences:

[LIV](2)-[SCT]-G-G-H-X(17,21)-D-D-[AST]-X-G-E-X(2)-D-K;
A-X(3)-P-G-L-X(3)-L-X(2)-G-X(13)-P-X(5)-H-X(3)-H;
[VIL]-L-[GSAT]-[VILFM]-E-[TS]-[TS]-C-D-[DE]; and
G-[LIV]-V-P-E-[LIV]-A-[AST]-R-X-H;

where,
the letters denote an amino acid in one letter code,
the square brackets denote a single amino acid,
the amino acids within the square brackets are alternatives,
X is any one amino acid residue, and
the numbers in the curved brackets refer to the number of residues at that position;

or

iv) [KR]-[GSAT]-X(4)-[FYWLH]-[DQNGK]-X-P-X-[LIVMFY]-X(3)-H-X(2)-[AG]-H-[LIVM]

where,

the letters denote an amino acid in one letter code,
the square brackets denote a single amino acid,
the amino acids within the square brackets are alternatives,
X is any one amino acid residue, and
the numbers in the curved brackets refer to the number of residues at that position.

2. A polypeptide or fragment according to claim 1 comprising all three of the sequences listed in iii).

3. A polypeptide containing any of the sequences set out in Figures 2a-2d.

4. A polypeptide according to any of claims 1-3 wherein said polypeptide is from *Borrella burgdorferi, Treponema pallidium, Synechocystis* sp. Strain PCC6803, *Helicobacter pylori, Arabidopsis thaliana, Haemophilus influenza, Mycobacterium tuberculosis, Mycobacterium leprae, Pasturella haemolytica, Mycoplasma genitalium, Mycoplasma pneumoniae, Streptococcus pneumoniae, Streptococcus pyogenes, Bacillus subtilis* or *Escherichia coli*.

5. A polypeptide according to any of claims 1-4 for use in a method of screening for agents with antibiotic activity.

6. An isolated polynucleotide encoding a polypeptide as defined in any of claims 1-4.

7. A vector comprising a transcriptional regulatory sequence and a nucleotide sequence encoding a polypeptide as defined in any of claims 1-4.

8. A host cell comprising a vector as claimed in claim 7 and a reporter gene whose activity is linked to the expression of the polypeptide according to any of claims 1-4.

9. A method of assaying compounds for activity against bacteria comprising:

i) providing a polypeptide according to the invention;
ii) contacting said polypeptide with an antagonist; and
iii) measuring for binding to said polypeptide.

10. A method of assaying compounds for activity against bacteria comprising:

i) expressing a polypeptide or fragment thereof according to any of claims 1-4 in a host cell;
ii) contacting said polypeptide with an antagonist; and
iii) measuring for inactivation of said polypeptide.

11. A method of assaying compounds for activity against bacteria comprising:

i) providing a polypeptide according to the invention;
ii) contacting said polypeptide with an antagonist; and
iii) measuring for cell death.

12. A method of assaying compounds for activity against bacteria comprising:

i) transfecting a host cell with a vector comprising a polynucleotide encoding a polypeptide as defined herein;
ii) allowing the host cell to express the polynucleotide;
iii) increasing the level of expression of the polypeptide as defined herein;
measuring for binding to said polypeptide; and
iv) assaying for increased resistance.

13. A method of assaying compounds for activity against bacteria comprising:

i) transfecting a host cell with a vector comprising a polynucleotide encoding a polypeptide as defined herein;
ii) allowing the host cell to express the polynucleotide;

iii) decreasing the level of expression of the polypeptide as defined herein; measuring for binding to said polypeptide; and

iv) assaying for increased sensitivity to an inhibitor.

14. A method of assaying compounds for activity against bacteria comprising:

i) generating a bacterial strain containing a reporter gene linked to the gene encoding a polypeptide according to the invention;

ii) contacting said strain with an antagonist; and

iii) measuring for induction or inhibition of said marker.

15. An antagonist of a polypeptide as defined in any of claims 1-4 identifiable by a method according to any of claims 9-14 for use in therapy.

16. Use of an antagonist of a polypeptide as defined in any of claims 1-4 identifiable by a method according to any of claims 9-14 for the manufacture of a medicament for the treatment of a bacterial infection.

17. A method of treatment which comprises administering to a patient an effective amount of an antagonist of a polypeptide as defined in any of claims 1-4 identifiable by a any of the methods according to claims 9-14.

(74)

FIG. 1

H. influenzae ----------
P. haemolytica ----------
E. coli ----------
M. leprae ----------
M. tuberculosis ----------
S. epidermidis ----------
B. subtilis ----------
S. pyogenes ----------
S. pneumoniae ----------
Synechocystis ----------
B. burgdorferi ----------
T. paladium ----------
M. genitalium ----------
M. pneumoniae ----------
A. thaliana MVRLFLTLSPAISRFNLYPGISILARNNNSLRLQKHHKLKTKTPTFSLISPSSSPNFQRT 60
H. pylori ----------

H. influenzae	----------MKILGIETSCDETGVAIYDEE---KGLIANQLYTQI	33
P. haemolytica	----------MRILGIETSCDETGVAIYDED---KGLVANQLYSQI	33
E. coli	----------MRVLGIETSCDETGIAIYDDE---KGLLANQLYSQV	33
M. leprae	----------MTISAVPGTIILAIETSCDETGVGIACLDDYGTVTLLADEVASSV	45
M. tuberculosis	----------MTTVLGIETSCDETGVGIARLDPDGTVTLLADEVASSV	38
S. epidermidis	----------	
B. subtilis	----------MSEQKDMYVLGIETSCDETAAAIVKNG---KEIISNVVASQI	39
S. pyogenes	----------MTDRYILAVESSCDETSVAILKNE---STLLSNVIASQV	36
S. pneumoniae	----------MKDRYILAFETSCDETSVAVLKND---DELLSNVIASQI	36
Synechocystis	----------MAIILAIETSCDETAVAIVNN-----RNVCSNVVSSQI	33
B. burgdorferi	----------MKVLGIETSCDDCCVAVVENG---IHILSNIKLNQT	33
T. paladium	----------KETKGGRRAVNVLGIETSCDETAVAIVKDG---THVCSNVVATQI	42
M. genitalium	----------MEQPLCVLGIETTCDDTGLSIVIDQ-----KIKSNIVISSA	36
M. pneumoniae	----------MEQPLCILGIETTCDDTSIGVITES-----KVQAHIVLSSA	36
A. thaliana	RFYSTETRISSLPYSENPNFDDNLVVLGIETSCDDTAAAVVSPFN---HLSS---SCRA	113
H. pylori	----------MILSIESSCDDSSLALTRIED---AQLIAHFKISQE	33

FIG. 1CONT'D

H. influenzae	ALHADYGGVVPELASRDHIRKTAPLIKAALEEANLT-ASDIDGIAYTSGPGLVGALLVGA	92
P. haemolytica	DMHADYGGVVPELASRDHIRKTLPLIQEALKEANLQ-PSDIDGIAYTAGPGLVGALLVGS	92
E. coli	KLHADYGGVVPELASRDHVRKTVPLIQAALKESGLT-AKDIDAVAYTAGPGLVGALLVGA	92
M. lepraae	DEQARFGGVVPEIASRAHLEALGPTIRCALAAAGLTGSAKPDVVAATIGPGLAGALLVGV	105
M. tuberculosis	DEHVRFGGVVPEIASRAHLEALGPAMRRALAAAGLK---QPDIVAATIGPGLAGALLVGV	95
S. epidermidis	----------MRFISKLVSAKVK-MEDIDAIAVTQGPGLIGALLIGI	36
B. subtilis	ESHKRFGGVVPEIASRHHVEQITLVIEEAFRKAGMT-YSDIDAIAVTEGPGLVGALLIGV	98
S. pyogenes	ESHKRFGGVVPEVASRHHVEVITTCFEDALQEAGIS-ASDLSAVAVTYGPGLVGALLVGL	95
S. pneumoniae	ESHKRFGGVVPEVASRHHVEVITACIEEALAEAGIT-EEDVTAVAVTYGPGLVGALLVGL	95
Synechocystis	QTHQIFGGVVPEVASRQHLLLINTCLDQALQASGLG-WPEIEAIAVTVAPGLAGALMVGV	92
B. burgdorferi	-EHKKYYGIVPEIASRLHTEAIMSVCIKALKKANTK-ISEIDLIAVTSRPGLIGSLIVGL	91
T. paladium	PFHAPYRGIVPELASRKHIEWILPTVKEALARAQLT-LADIDGIAVTHAPGLTGSLLVGL	101
M. genitalium	NLHVKTGGVVPEIAARCHEQN----LFKAIRDLNFE-IRDLSHIAYACNPGLAGCLHVGA	91
M. pneumoniae	KLHAQTGGVVPEVAARSHEQN----LLKALQQSGVV-LEQITHIAYAANPGLPGCLHVGA	91
A. thaliana	ELLVQYGGVAPKQAEEAHSRVIDKVVQDALDKANLT-EKDLSAVAVTIGPGLSLCLRVGV	172
H. pylori	KHHSSYGGVVPELASRLHAEN-LPLLLERIKISLNKDFSKIKAIAITNQPGLSVTLIEGL	92

FIG. 1CONT'D

H. influenzae	TIARSLAYAWNVPAIGVHHMEGHLLAPMLDDNS--PHFPFVALLVSGGHTQLVRVDGVGK	150
P. haemolytica	TIARSLAYAWNVPALGVHHMEGHLLAPMLEENA--PEFPFVALLISGGHTQLVKVDGVGQ	150
E. coli	TVGRSLAFAWDVPAIPVHHMEGHLLAPMLEDNP--PEFPFVALLVCGGHTQLISVTGIGQ	150
M. leprae	AAAKAYSAAWGVPFYAVNHLGGHLAADVYEHG---PLPECVALLVSGGHTHLLQVRSLGA	162
M. tuberculosis	AAAKAYSAAWGVPFYAVNHLGGHLAADVYEHG---PLPECVALLVSGGHTHLLHVRSLGE	152
S. epidermidis	NAAKALAFAYDKPIIPVHHIAGHIYANHLEQP---LTFPLMSLIVSGGHTELVYMKNHLD	93
B. subtillis	NAAKALSFAYNIPLVGVHHIAGHIYANRLVED---IVFPALALVVSGGHTELVYMKEHGS	155
S. pyogenes	AAAKAFAWANHLPLIPVNHMAGHLMAAREQKP---LVYPLIALLVSGGHTELVYVPEPGD	152
S. pneumoniae	SAAKAFAWAHGLPLIPVNHMAGHLMAAQSVEP---LEFPLLALLVSGGHTELVYVSEAGD	152
Synechocystis	TAAKTLAMVHQKPFLGVHHLEGHIYASYLSQPD--LQPPFLCLLVSGGHTSLIHVKGCGD	150
B. burgdorferi	NFAKGLAISLKKPIICIDHILGHLYAPLMHSK---IEYPFISLLLSGGHTLIAKQKNFDD	148
T. paladium	TFAKTLAWSMHLPFIAVNHLHAHFCAAHVEHD---LAYPYVGLLASGGHALVCVVHDFDQ	158
M. genitalium	TFARSLSFLLDKPLLPINHLYAHIFSCLIDQDLNKLQLPALGLVISGGHTAIYLVKSFYE	151
M. pneumoniae	TFARSLSFLLDKPLLPINHLYAHIFSALIDQDINQLKLPALGLVVSGGHTAIYLIKSLFD	151
A. thaliana	RKARRVAGNFSLPIVGVHHMEAHALVARLVEQ--ELSFPFMALLISGGHNLLVLAHKLGQ	230
H. pylori	MMAKALSLSLNLPLILEDHLRGHVYSLFINEKQ--TCMPLSVLLVSGGHSLILEARDYEN	150

FIG. 1CONT'D

H. influenzae	-YEVIGESIDDAAGEAFDKTAKLLGLDYP--GGAALSRLAEKGTPN--RFTFPRPMTDRA	205
P. haemolytica	-YELLGESIDDAAGEAFDKTGKLLGLDYP--AGVAMSKLAESGTPN--RFKFPRPMTDRP	205
E. coli	-YELLGESIDDAAGEAFDKTAKLLGLDYP--GGPLLSKMAAQGTAG--RFVFPRPMTDRP	205
M. lepraе	PIVELGSTVDDAAGEAYDKVARLLGLGYP--GGKVLDDLARTGDRD--AIVFPRGMTGPA	218
M. tuberculosis	PIIELGSTVDDAAGEAYDKVARLLGLGYP--GGKALDDLARTGDRD--AIVFPRGMSGPA	208
S. epidermidis	-FEVIGETRDDAVGEAYDKVARTINLPYP--GGPHIDRLAAKGKD---VYDFPRVWLEKD	147
B. subtilis	-FEVIGETLDDAAGEAYDKVARTMGLPYP--GGPQIDKLAEKGND---NIPLPRAWLEEG	209
S. pyogenes	-YHIIGETRDDAVGEAYDKVGRVMGLTYP--AGREIDQLAHKGQD---TYHFPRAMITED	206
S. pneumoniae	-YKIVGETRDDAVGEAYDKVGRVMGLTYP--AGREIDELVHQGQD---IYDFPRAMIKED	206
Synechocystis	-YRQLGTTRDDAAGEAFDKVARLLDLGYP--GGPAIDRAAKQGDPG--TFKLPEGKISLP	205
B. burgdorferi	-VEILGRTLDDACGEAFDKVAKHYDMGFP--GGPNIEQISKNGDEN--TFQFPVTTFKKK	203
T. paladium	-VEALGATIDDAPGEAFDKVAAFYGFGYP--GGKVIETLAEQGDAR--AARFPLPHFHGK	213
M. genitalium	-LELIAETSDDAIGEVYDKIGRAMGFDYP--AGSKIDSLFNKELVK--PHYFFKPSTKWT	206
M. pneumoniae	-LELIAETSDDAIGEVYDKVGRAMGFPYP--AGPQLDSLFQPELVK--SHYFFRPSTKWT	206
A. thaliana	-YTQLGTTVDDAIGEAFDKTAKWLGLDMHRSGGPAVEELALEGDAK--SVKFNVPMKYHK	287
H. pylori	-IKIVATSLDDSFGESFDKVSKMLDLGYP--GGPIVEKLALDYRHPNEPLMFPIPLKNSP	207

FIG. 1CONT'D

H. influenzae	G-----LDFSFSGLKTFAANTVNQAIKNEGE-LIEQTKADIAYAFQDAVVDTLAIKCKRA	259
P. haemolytica	G-----LDFSFSGLKTFAANTIKANLNENGE-LDEQTKCDIAHAFQQAVVDTILIKCKRA	259
E. coli	G-----LDFSFSGLKTFAANTIR----DNG--TDDQTRADIARAFEDAVVDTLMIKCKRA	254
M. lepraе	D---DLNAFSFSGLKTAVARYVESH---------PDALPADVAAGFQEAVADVLTMKAVRA	267
M. tuberculosis	D---DRYAFSFSGLKTAVARYVESHA--AD---PGFRTADIAAGFQEAVADVLTMKAVRA	260
S. epidermidis	S-----YDFSFSGLKSAVINKLHNLR-QKN---IEIVAEDVATSFQNSVVEVLTYKAIHA	198
B. subtilis	S-----YNFSFSGLKSAVINTLHNAS-QKG---QEIAPEDLSASFQNSVIDVLVTKTARA	260
S. pyogenes	H-----LEFSFSGLKSAFINLHHNAK-QKG---DELILEDLCASFQAAVLDILLAKTKKA	257
S. pneumoniae	N-----LEFSFSGLKSAFINLHHNAE-QKG---ESLSTEDLCASFQAAVMDILMAKTKKA	257
Synechocystis	QGGYHPYDSSFSGLKTAMLRLTQELK-QSS---APLPVDDLAASFQDTVARSLTKKTIQC	261
B. burgdorferi	E---NWYDFSYSGLKTACIHQLEKFK-SKD---NPTTKNNIAASFQKAAFENLITPLKRA	256
T. paladium	G---HRYDVSYSGLKTAVIHQLDHFW-NKE---YERTAQNIAAAFQACAINILLRPLARA	266
M. genitalium	K---------FSYSGLKSQCLNKIKQIS----ANKTRIDWSELASNFQATIIDHYIDHVKNA	255
M. pneumoniae	K---------FSYSGLKSQCFTKIKQLRERKGFNPQTHDWNEFASNFQATIIDHYINHVKDA	259
A. thaliana	D-----CNFSYAGLKTQVRLAIEAK----------EIRNRADIAASFQRVAVLHLEEKCERA	334
H. pylori	N-----LAFSFSGLKNAVRLEVEKNAPNLN----EAIKQKIGYHFQSAAIEHLIQQTKRY	258

FIG. 1CONT'D

H. influenzae	LK-----ETGYKRLVIAGGVSANKKLRETLAHLMQNLG-GEVFYPQPQFCTDNGAMIAYT	313
P. haemolytica	LE-----QTGYKRLVMAGGVSANKQLRADLAEMMKKLK-GEVFYPRPQFCTDNGAMIAYT	313
E. coli	LD-----QTGFKRLVMAGGVSANRTLRAKLAEMMKKRR-GEVFYARPEFCTDNGAMIAYA	308
M. lepraе	AT-----GLGVSTLLIVGGVAANSRLR-ELAAQRCAAAGLMLRIPGPRFCTDNGAMIAAF	321
M. tuberculosis	AT-----ALGVSTLLIAGGVAANSRLR-ELATQRCGEAGRTLRIPSPRLCTDNGAMIAAF	314
S. epidermidis	CK-----TYNVNRLIVAGGVASNKGLRNALS-EACKKEGIHLTIPSPVLCTDNAAMIGAA	252
B. subtilis	AK-----EYDVKQVLLAGGVAANRGLRAALEKEFAQHEGITLVIPPLALCTDNAAMIAAA	315
S. pyogenes	LS-----RYPAKMLVVAGGVAANQGLRDRLAQEITH---IEVVIPKLRLCGDNAGMIALA	309
S. pneumoniae	LE-----KYPVKTLVVAGGVAANKGLRERLAAEITD---VKVIIPPLRLCGDNAGMIAYA	309
Synechocystis	VL-----DHGLTTITVGGGVAANSRLRYHLQTAAQEHQ-LQVFFPPLKFCTDNAAMIACA	315
B. burgdorferi	IK-----DTQINKLVIAGGVASNLYLREKIDKLK-----IQTYYPPLDLCTDNGAMIAGL	306
T. paladium	LQ-----DTGLPTAVVCGGVAANSLLRKSVADWKH----ARCVFPSREYCTDNAVMVAAL	317
M. genitalium	IK-----KFAPKMLLVGGGVSANSYLSNRISTLN-----LPFLIADSKYTSDNGAMIGFY	305
M. pneumoniae	IQ-----QHQPQMLLLGGGVSANKYLREQVTQLQ-----LPYLIAPLKYTSDNGAMIGFY	309
A. thaliana	IDWALELEPSIKHMVISGGVASNKYVRLRLNNIVENKN-LKLVCPPPSLCTDNGVMVAWT	393
H. pylori	FK-----IKRPKIFGIVGGASQNLALRKAFENLCDAFD-CKLVLAPLEFCSDNAAMIGRS	312

FIG. 1CONT'D

H. influenzae	GFLRLKQGQH----------SDLA-IDVKPRWAMAELPAI----------	342
P. haemolytica	GFLRLKTMN----------KPT----------	325
E. coli	GMVRFKAGAT----------ADLG-VSVRPRWPLAELPAA----------	337
M. lepraе	AAHLLAAAAPP----------SPLD-VPSDPGLPVVKRQIN----------	351
M. tuberculosis	AAQLVAAGAPP----------SPLD-VPSDPGLPVMQGQVR----------	344
S. epidermidis	GYYLYQAGLR----------GDLA-LNGQNNIDIETFSV----------	280
B. subtilis	GTIAFEKGIR----------GAYD-MNGQPGLELTSYQSLTR----------	346
S. pyogenes	AAIEYDKQHF----------ANMS-LNAKPSLAFDQFPDSFVIN----------	342
S. pneumoniae	SVSEWNKENF----------AGWD-LNAKPSLAFDTME----------	336
Synechocystis	AADHFQNGDR----------SPLT-LGVQSRLSVEQVSQLYERN----------	348
B. burgdorferi	GFNMYLKYGE----------SPIE-IDANSRIENYKNQYRGKNNEKNFSNA----------	346
T. paladium	GYRYLIRGDR----------SFYG-VTERSRIAHFS-KRGGDRLAAQRSAASQPLF----------	361
M. genitalium	ASLLINGDKN----------	315
M. pneumoniae	ANLLINGKNN----------	319
A. thaliana	GLEHFRVGRYDPPPPATEPEDYV-YDLRPRWPLGEEYAKGRSEARSMRTARIHPSLTSII	452
H. pylori	SLEAYQKKRF----------VPLEKANISPRTLLKSFE----------	340

FIG. 1CONT'D

H. influenzae	----------	
P. haemolytica	----------	
E. coli	----------	
M. leprae	----------	
M. tuberculosis	----------	
S. epidermidis	----------	
B. subtilis	----------	
S. pyogenes	----------	
S. pneumoniae	----------	
Synechocystis	----------	
B. burgdorferi	----------	
T. paladium	----------	
M. genitalium	----------	
M. pneumoniae	----------	
A. thaliana	RADSLQQQTQT	463
H. pylori	----------	

FIG. 1CONT'D

MOTIF 1

Sequence name	Start	MOTIF
E. coli	128	APMLEDNPPE FPFVALLVCGGHTQLISVTGIGQYELLGES IDDAAGEAFDKTAKLLGLDYPGG PLLSKMAAQG
H. influenzae	128	APMLDDNSPH FPFVALLVSGGHTQLVRVDGVGKYEVIGES IDDAAGEAFDKTAKLLGLDYPGG AALSRLAEKG
P. haemolytica	128	APMLEENAPE FPFVALLISGGHTQLVKVDGVGQYELLGES IDDAAGEAFDKTGKLLGLDYPAG VAMSKLAESG
S. epidermidis	71	YANHLEQPLT FPLMSLIVSGGHTELVYMKNHLDFEVIGET RDDAVGEAYDKVARTINLPYPGG PHIDRLAAKG
B. subtilis	133	YANRLVEDIV FPALALVVSGGHTELVYMKEHGSFEVIGET LDDAAGEAYDKVARTMGLPYPGG PQIDKLAEKG
H. pylori	128	SLFINEKQTC MPLSVLLVSGGHSLILEARDYENIKIVATS LDDSFGESFDKVSKMLDLGYPGG PIVEKLALDY
M. genitalium	129	LIDQDLNKLQ LPALGLVISGGHTAIYLVKSFYELELIAET SDDAIGEVYDKIGRAMGFDYPAG SKIDSLFNKE
M. pneumoniae	129	LIDQDINQLK LPALGLVVSGGHTAIYLIKSLFDLELIAET SDDAIGEVYDKVGRAMGFPYPAG PQLDSLFQPE
Synechocystis	128	ASYLSQPDLQ PPFLCLLVSGGHTSLIHVKGCGDYRQLGTT RDDAAGEAFDKVARLLDLGYPGG PAIDRAAKQG
B. burgdorferi	126	YAPLMHSKIE YPFISLLLSGGHTLIAKQKNFDDVEILGRT LDDACGEAFDKVAKHYDMGFPGG PNIEQISKNG
M. leprae	141	ADVYEHGPLP ECVALLVSGGHTHLLQVRSLGAPIVELGST VDDAAGEAYDKVARLLGLGYPGG KVLDDLARTG
M. tuberculosis	131	ADVYEHGPLP ECVALLVSGGHTHLLHVRSLGEPIIELGST VDDAAGEAYDKVARLLGLGYPGG KALDDLARTG
S. pyogenes	130	MAAREQKPLV YPLIALLVSGGHTELVYVPEPGDYHIIGET RDDAVGEAYDKVGRVMGLTYPAG REIDQLAHKG
T. paladium	136	CAAHVEHDLA YPYVGLLASGGHALVCVVHDFDQVEALGAT IDDAPGEAFDKVAAFYGFGYPGG KVIETLAEQG
S. pneumoniae	130	MAAQSVEPLE FPLLALLVSGGHTELVYVSEAGDYKIVGET RDDAVGEAYDKVGRVMGLTYPAG REIDELVHQG
A. thaliana	208	VARLVEQELS FPFMALLISGGHNLLVLAHKLGQYTQLGTT VDDAIGEAFDKTAKWLGLDMHRS GGPAVEELAL

FIG. 2a

MOTIF 2

Sequence name	Start	MOTIF
E. coli	72	ALKESGLTAK DIDAVAYTAGPGLVGALLVGATVGRSLAFA WDVPAIPVHHMEGHLLA PMLEDNPPEF
H. influenzae	72	ALEEANLTAS DIDGIAYTSGPGLVGALLVGATIARSLAYA WNVPAIGVHHMEGHLLA PMLDDNSPHF
P. haemolytica	72	ALKEANLQPS DIDGIAYTAGPGLVGALLVGSTIARSLAYA WNVPALGVHHMEGHLLA PMLEENAPEF
S. epidermidis	16	KLVSAKVKME DIDAIAVTQGPGLIGALLIGINAAKALAFA YDKPIIPVHHIAGHIYA NHLEQPLTFP
B. subtilis	78	AFRKAGMTYS DIDAIAVTEGPGLVGALLIGVNAAKALSFA YNIPLVGVHHIAGHIYA NRLVEDIVFP
H. pylori	72	IKISLNKDFS KIKAIAITNQPGLSVTLIEGLMMAKALSLS LNLPLILEDHLRGHVYS LFINEKQTCM
M. genitalium	71	AIRDLNFEIR DLSHIAYACNPGLAGCLHVGATFARSLSFL LDKPLLPINHLYAHIFS CLIDQDLNKL
M. pneumoniae	71	ALQQSGVVLE QITHIAYAANPGLPGCLHVGATFARSLSFL LDKPLLPINHLYAHIFS ALIDQDINQL
Synechocystis	72	ALQASGLGWP EIEAIAVTVAPGLAGALMVGVTAAKTLAMV HQKPFLGVHHLEGHIYA SYLSQPDLQP
B. burgdorferi	71	ALKKANTKIS EIDLIAVTSRPGLIGSLIVGLNFAKGLAIS LKKPIICIDHILGHLYA PLMHSKIEYP
M. leprae	85	LAAAGLTGSA KPDVVAATIGPGLAGALLVGVAAAKAYSAA WGVPFYAVNHLGGHLAA DVYEHGPLPE
M. tuberculosis	75	RRALAAAGLK QPDIVAATIGPGLAGALLVGVAAAKAYSAA WGVPFYAVNHLGGHLAA DVYEHGPLPE
S. pyogenes	75	ALQEAGISAS DLSAVAVTYGPGLVGALLVGLAAAKAFAWA NHLPLIPVNHMAGHLMA AREQKPLVYP
T. paladium	81	ALARAQLTLA DIDGIAVTHAPGLTGSLLVGLTFAKTLAWS MHLPFIAVNHLHAHFCA AHVEHDLAYP
S. pneumoniae	75	ALAEAGITEE DVTAVAVTYGPGLVGALLVGLSAAKAFAWA HGLPLIPVNHMAGHLMA AQSVEPLEFP
A. thaliana	152	ALDKANLTEK DLSAVAVTIGPGLSLCLRVGVRKARRVAGN FSLPIVGVHHMEAHALV ARLVEQELSF

FIG. 2b

MOTIF 3

Sequence name	Start	Site	
E. coli	3	MR	VLGIETSCDET GIAIYDDEKG
H. influenzae	3	MK	ILGIETSCDET GVAIYDEEKG
P. haemolytica	3	MR	ILGIETSCDET GVAIYDEDKG
B. subtilis	9	MSEQKDMY	VLGIETSCDET AAAIVKNGKE
H. pylori	2	M	ILSIESSCDDS SLALTRIEDA
M. genitalium	7	MEQPLC	VLGIETTCDDT GLSIVIDQKI
M. pneumoniae	7	MEQPLC	ILGIETTCDDT SIGVITESKV
Synechocystis	4	MAI	ILAIETSCDET AVAIVNNRNV
B. burgdorferi	3	MK	VLGIETSCDDC CVAVVENGIH
M. leprae	11	MTISAVPGTI	ILAIETSCDET GVGIACLDDY
M. tuberculosis	4	MTT	VLGIETSCDET GVGIARLDPD
S. pyogenes	6	MTDRY	ILAVESSCDET SVAILKNEST
T. paladium	12	ETKGGRRAVN	VLGIETSCDET AVAIVKDGTH
S. pneumoniae	6	MKDRY	ILAFETSCDET SVAVLKNDDE
A. thaliana	86	ENPNFDDNLV	VLGIETSCDDT AAAVVSPFNH

FIG. 2c

MOTIF 4 width = 10

Sequence name	Start	Site	
E. coli	31	EKGLLANQLY	SQVKLHADYGGVVPELASRDH VRKTVPLI QA
H. influenzae	31	EKGLIANQLY	TQIALHADYGGVVPELASRDH IRKTAPLI KA
P. haemolytica	31	DKGLVANQLY	SQIDMHADYGGVVPELASRDH IRKTLPLI QE
B. subtilis	37	GKEIISNVVA	SQIESHKRFGGVVPEIASRHH VEQITLVI EE
H. pylori	31	DAQLIAHFKI	SQEKHHSSYGGVVPELASRLH AENLPLLL ER
M. genitalium	34	DQKIKSNIVI	SSANLHVKTGGVVPEIAARCH EQNLFKAI RD
M. pneumoniae	34	ESKVQAHIVL	SSAKLHAQTGGVVPEVAARSH EQNLLKAL QQ
Synechocystis	31	NRNVCSNVVS	SQIQTHQIFGGVVPEVASRQH LLLINTCL DQ
B. burgdorferi	30	NGIHILSNIK	LNQTEHKKYYGIVPEIASRLH TEAIMSVC IK
M. leprae	43	TVTLLADEVA	SSVDEQARFGGVVPEIASRAH LEALGPTI RC
M. tuberculosis	36	TVTLLADEVA	SSVDEHVRFGGVVPEIASRAH LEALGPAM RR
S. pyogenes	34	ESTLLSNVIA	SQVESHKRFGGVVPEVASRHH VEVITTCF ED
T. paladium	40	GTHVCSNVVA	TQIPFHAPYRGIVPELASRKH IEWILPTV KE
S. pneumoniae	34	DDELLSNVIA	SQIESHKRFGGVVPEVASRHH VEVITACI EE
A. thaliana	111	VSPFNHLSSS	CRAELLVQYGGVAPKQAEEAH SRVIDKVV QD

FIG. 2d

Prosite pattern 1
[liv](2)-[sct]-g-g-h-x(17,21)-d-d-[AST]-x-g-e-x(2)-d-k

Prosite pattern 2
a-x(3)-p-g-l-x(3)-l-x(2)-g-x(13)-p-x(5)-h-x(3)-h

Prosite pattern 3
[VIL]-l-[GSAT]-[VILFM]-e-[TS]-[TS]-c-d-[DE]

Prosite pattern 4
g-[LIV]-v-p-e-[LIV]-a-[AST]-r-x-h

Prosite pattern PS01016 - Glycoprotease family signature

[KR]-[GSAT]-x(4)-[FYWLH]-[DQNGK]-x-P-x-[LIVMFY]-x(3)-H-x(2)-[AG]-H-[LIVM]

FIG. 3

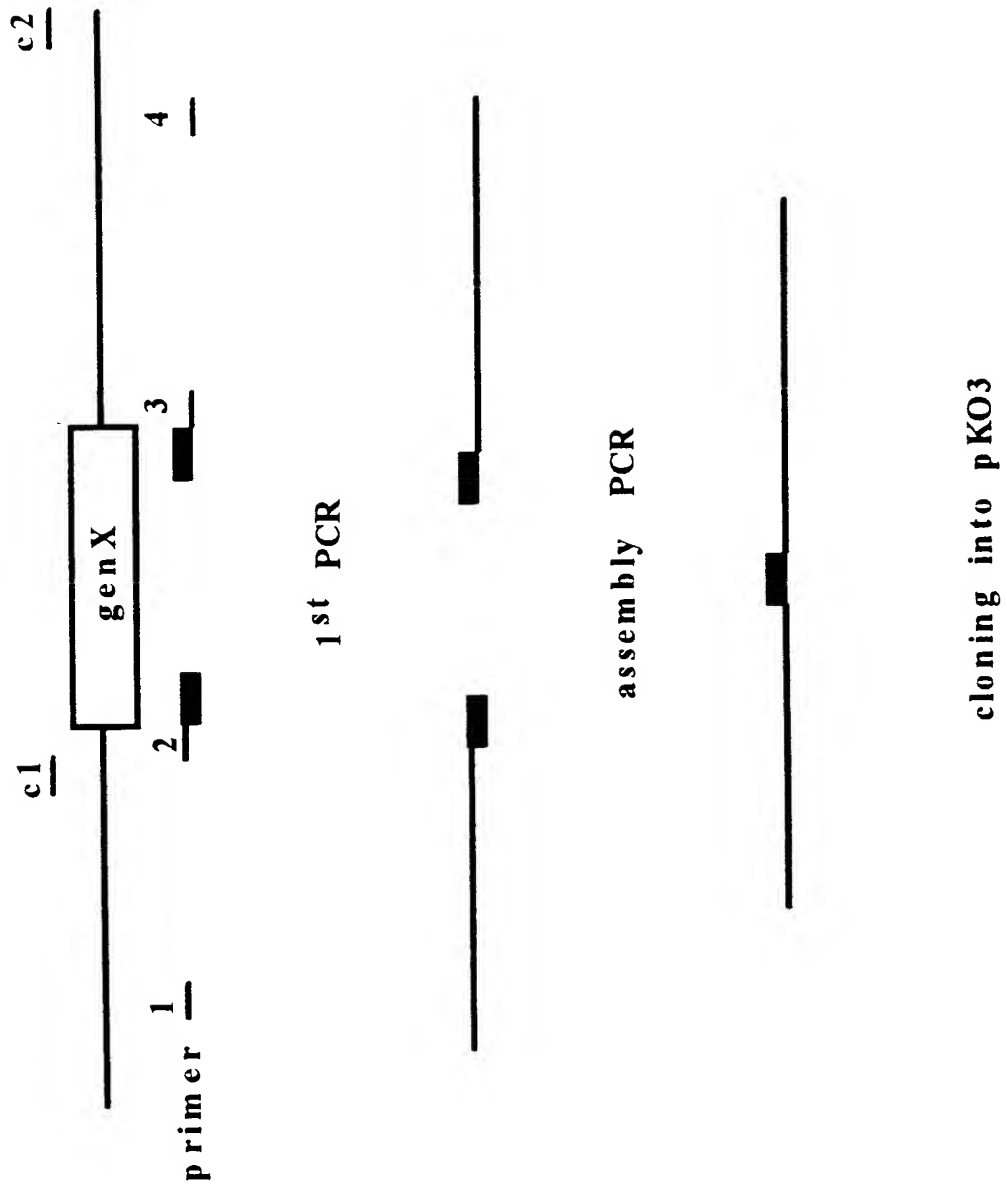
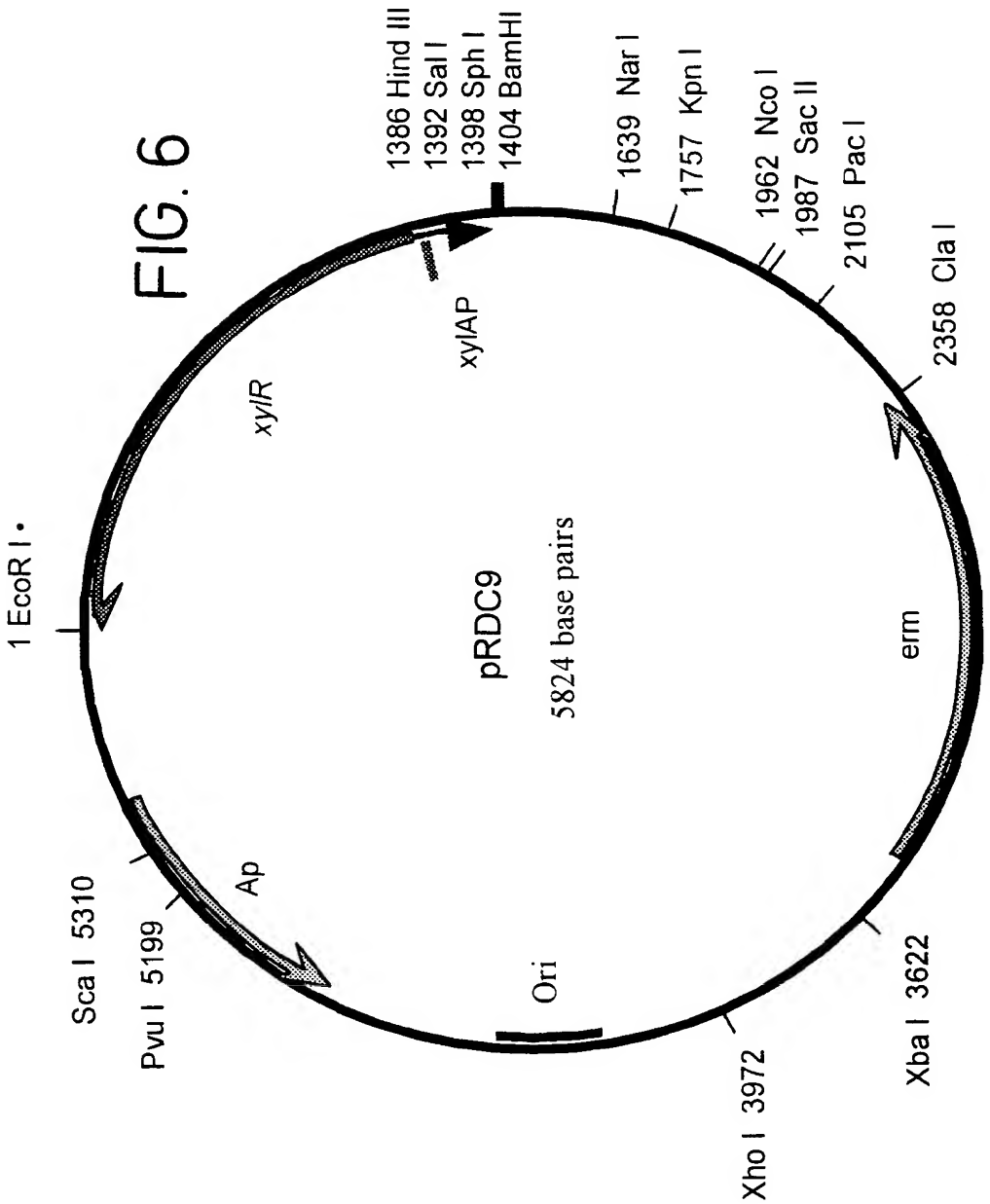
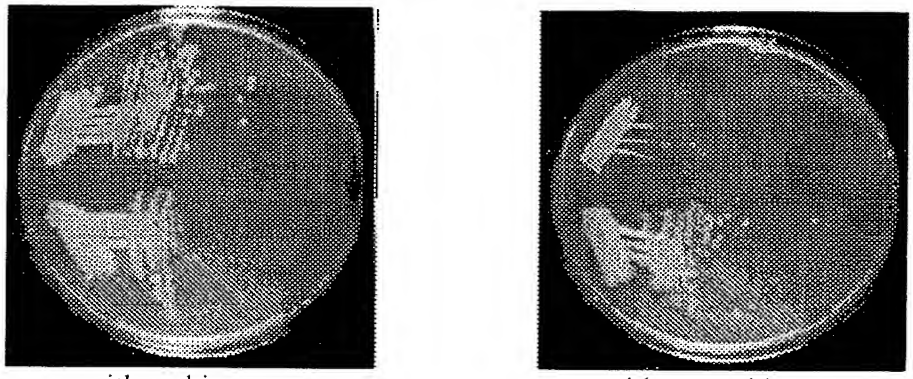

c2
4
3
genX
2
c1
1
primer
1st PCR
assembly PCR
cloning into pKO3

FIG. 4

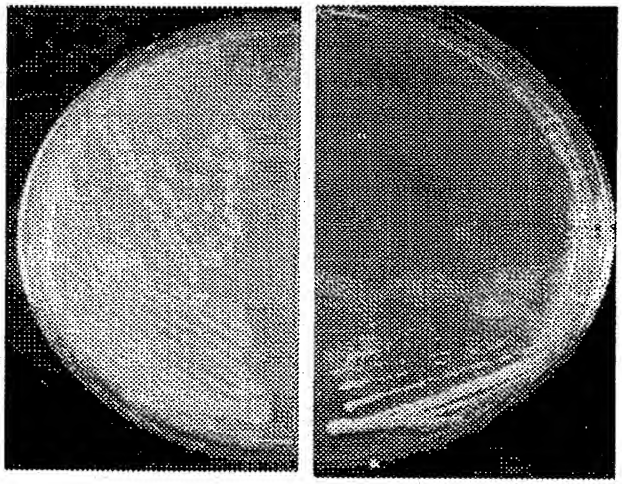
FIG. 6


1 EcoR I
Sca I 5310
Pvu I 5199
Ap
Ori
Xho I 3972
Xba I 3622
erm
2358 Cla I
2105 Pac I
1987 Sac II
1962 Nco I
1757 Kpn I
1639 Nar I
1404 BamHI
1398 Sph I
1392 Sal I
1386 Hind III
xylAP
xylR
pRDC9
5824 base pairs



with arabinose without arabinose

FIG. 5



with xylose without xylose

FIG. 7

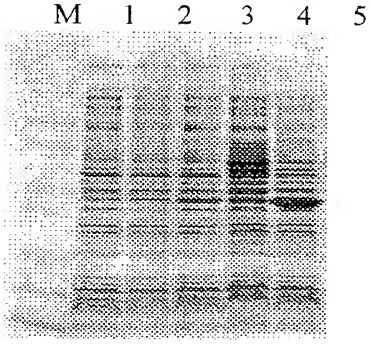


FIG. 8